DOCUMENT RESUME

ED 457 729 HE 034 378

AUTHOR	Gonzalez, Julie M. Byers; DesJardins, Stephen L.
TITLE	Artificial Neural Networks: A New Approach for Predicting Application Behavior. AIR 2001 Annual Forum Paper.
PUB DATE	2001-06-05
NOTE	38p.; Paper presented at the Annual Meeting of the Association for Institutional Research (41st, Long Beach, CA, June 3-6, 2001). SPSS, Inc. provided the software for this study.
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Cohort Analysis; *College Applicants; Comparative Analysis; Higher Education; Models; *Prediction; *Statistical Analysis
IDENTIFIERS	*Neural Networks

ABSTRACT

This paper examines how predictive modeling can be used to study application behavior. A relatively new technique, artificial neural networks (ANNs), was applied to help predict which students were likely to get into a large Research I university. Data were obtained from a university in Iowa. Two cohorts were used, each containing approximately 20,000 records. The results of the new techniques were compared to those from the traditional analysis tool, logistic regression modeling. ANNs have several advantages over traditional statistical methods. ANNs do not require knowledge of the functional relationship between the independent variables and the dependent variables to estimate the model. Unlike traditional statistical techniques, ANNs learn from examples with a small number of prior assumptions about structural relationships. Other advantages and disadvantages are discussed. The addition of artificial intelligence models is an exciting new area that may be used to explore the complex processes in educational institutions. (Contains 3 figures, 7 tables, and 48 references.) (SLD)

Artificial Neural Networks: A New Approach for Predicting Application Behavior

Julie M. Byers González
Ph.D. Candidate
Educational Policy and Leadership Studies
N491 Lindquist Center
Iowa City, Iowa 52242
julie-gonzalez@uiowa.edu

Stephen L. DesJardins
Assistant Professor
Educational Policy and Leadership Studies
N491 Lindquist Center
Iowa City, Iowa 52242
stephen-desjardins@uiowa.edu

June 5th, 2001

Association for Institutional Research (AIR) 41st Annual Forum

Long Beach, California

Correspondence may be directed to: Stephen L. DesJardins at the above address.

Acknowledgement: We would like to thank SPSS, Inc. for providing the Clementine software used in this study.

Running Head: Artificial Neural Networks

Artificial Neural Networks: A New Approach for Predicting Application Behavior

Abstract

In this paper we examine how predictive modeling can be used to study application behavior. We apply a relatively new technique, artificial neural networks, to help us predict which students are likely to apply to a large Research I institution. We compare the results of these new techniques to the traditional analysis tool, logistic regression modeling. The addition of artificial intelligence models is an exciting new area and hopefully this paper will encourage other institutional researchers to use this technique to explore the complex processes found in our educational institutions.

Introduction

Research on college choice has been primarily grounded in economic, sociological, and combined models (Kohn, Manski, & Mundel, 1976; Litten, 1982). Econometric models presume that students aim to maximize their utility or expected benefit from their choice of institution (Bruggink & Gambhir, 1996; Ganderton, 1992; Hossler, Braxton, & Coopersmith, 1989; Paulsen, 1990; Welki & Navratil, 1987). Moreover, these models assume that students have knowledge of the colleges' characteristics, costs, benefits, and will behave in a manner that will tend to maximize benefits (Hamrick & Hossler, 1996). Sociological or status attainment models focus on college aspirations. These models are derived from the general status attainment literature that focuses on the identification and interrelationship of factors and characteristics influencing college aspirations (Hossler et al., 1989; McDonough, 1994; Paulsen, 1990). Combined models, on the other hand, are deduced from both econometric and sociological models. These models seek to account for diverse economic and social forces that influence decision- making.

While extensive college choice literature exists, there is a need for more theory building and additional empirical research in order to understand better the complex process of college choice. Additional research is especially needed in the area of student application behavior because the college choice literature has typically focused on students' enrollment decisions (Manski & Wise, 1983). Few studies have examined college application behavior (for exceptions see Weiler, 1994; -------------,1999; and Goyette, 1999).

This study helps to address the relative dearth of research in this area by utilizing artificial neural network (ANN) techniques to predict college application behavior at a Research I institution in Iowa. We also compare how the results of the ANN model compare to traditional statistical methods used to study student application behavior. We also document how predictive modeling contributed to institutional recruitment efforts at the study institution. This study is unique in that we know of no other study that utilized artificial neural network modeling as the primary analytical tool to investigate student application behavior.

Literature Review

Student Choice Literature

Litten (1982) portrays the college choice process as a funnel that is partitioned into three stages. Sociological background and personal characteristics comprise the first stage while institutional and economic variables are important during the second and third stages of the model (Bateman & Spruill, 1996). The funnel approach presumes that college choice involves a process of elimination by which the ultimate decision to enroll at a particular institution is the result of a sequential process. Litten determined that high school characteristics, student performance, and high school curriculum tended to affect student aspirations to attend a particular college. Moreover, socioeconomic status, personal characteristics, and the economic, political, and cultural climate also affect a student's propensity to attend a specific institution. The influence of parents, peers, and close friends was also determined to be substantial at the information gathering stage. (Hossler et al., 1989).

In Hossler and Gallagher's (1987) Three Phase Development Model, the first stage of college choice is referred to as the predisposition stage. The predisposition stage involves the development of educational aspirations and it is at this stage that student decision-making processes begin to take shape. While the decision-making process of students during the predisposition phase is not well understood, research demonstrates a positive relationship between the quality of the high school curriculum and eventual college enrollment (Hearn, 1984; Kolstad, 1979). The second stage is known as the search stage. At this stage students search for an appropriate institution and institutions search for the appropriate applicant fit. Insight into student decision-making processes at the predisposition and search stage may prove instrumental in understanding why certain students decide to apply and eventually enroll in a particular college or university. The final stage is the college choice, where students evaluate the institutions in their "choice set" and choose one to enroll in.

Factors Influencing College Application

Timing and accurate information is especially important during the application process. While information may come from a variety of sources, parents are the most influential source (Chapman, 1981). Parents who have completed a post-secondary education will be more capable of giving their children good advice, and this relationship is especially important since children typically depend upon their parents for guidance and support. Greater accessibility to social capital and reliable information and resources will tend to more effectively guide students through the college application process. McLanahan and Sandefur (1994) note that intact families provide the social capital necessary for students to more effectively apply to college, where as broken families often lack this type of support. Researchers have found that Asian students are more likely than their Caucasian counterparts to come from intact families (Kitano & Daniels, 1988; Min, 1988) and may be more likely possess the social capital necessary to navigate the college application process (Goyette, 1999).

Higher levels of socioeconomic status commonly permit families to supply better material resources to their children in the form of conducting more sophisticated college searches, collecting information on college application procedures, and providing the financial resources for ACT or SAT preparatory classes. There are a few major distinctions in the application practices of students from different socioeconomic backgrounds. For example, McDonough (1994) noted that students from higher socioeconomic backgrounds filed an average of ten applications while students from lower socioeconomic backgrounds filed an average of two to three applications. One reason for lower application rates from students from lower socioeconomic backgrounds may be the cost of application fees, which often range from $15 to $50 (McDonough, 1994).

Research also indicates that people inquire about information from individuals of comparable socioeconomic backgrounds (Stanton-Salazar & Dornbusch, 1995). Stanton-Salazar and Dornbusch maintain that the higher the socioeconomic background of the person supplying the information with regard to college choice, the greater the chances are of that information being reliable, accurate, and pertinent. For these reasons, students of higher socioeconomic backgrounds are likely to receive more appropriate information with regard to college application procedures than their lower SES counterparts.

The educational attainment of one's parents is also positively related to a student's educational expectations (Goyette, 1999). Goyette determined that higher educational expectations are thought to explain the higher college application rates among Asian American groups.

With regard to institutional recruitment efforts, Freeman (1984) found that actions such as inviting students to a banquet, personal letters from the college president, or special certificates were influential during the choice phase for all types of students. Research also indicates that non-aid based courtship procedures, such as campus visits are highly effective (Freeman, 1984). For example, approximately 40% of seniors who make a campus visit eventually apply to the institution (Dehne, 1994).

However, many institutions are still uncertain that marketing and recruitment activities really work (----------, 1999). Moreover, the college choice literature is not complete, especially with regard to the analysis of the factors affecting student choice at different types of institutions. Further, many institutions lack information regarding the factors affecting prospective students' college choice propensities (-----------, 1999). Additionally, a variety of issues such as shifting demographic patterns, increased student applications, and a student buyers' market suggest that we need to reexamine college application behavior.

Current Issues in the College Application Literature

Research indicates that the 1980's and 1990's was a period in which the number of high school graduates was below historical trends. From 1981 to 1986 the number of high school graduates dropped 14% across the nation (Bryant & Crockett, 1993; Melia & Goodman, 1988; Rainsford, 1985). This trend was followed by an additional decrease of

200,000 high school graduates by 1992 (Bryant & Crockett, 1993; Melia & Goodman, 1988; Rainsford, 1985). While the 1990's was a period of relatively few high school graduates, colleges and universities ironically encountered above average applications (McDonough, 1994) because students tended to file more applications than in the past (Dey, Astin, & Korn, 1991). This may also complicate our understanding of the college choice process because increased student applications do not necessarily translate into increased enrollments. Therefore, an increase in the number of applicants may not be a good predictor of future enrollments (Sanoff, 1994).

Today's college prospects are also inclined to apply to more prestigious institutions (Shea, 1994) and are likely to pay enrollment deposits at multiple institutions (Hossler, Schmit, & Vesper, 1999). High ability students, who are increasingly the focus of many institutional recruitment efforts, are likely to apply to and research a greater number of colleges and universities (Galotti & Mark, 1994) and have a greater propensity to conduct more efficient and better refined searches (Ihlanfeldt, 1980; Litten, Sullivan, & Brodigan, 1983). Students also begin the college search process much earlier than they used to. In 1975, Lewis and Morrison determined that only 10% of high school seniors began to inquire about college information during October of their senior year (cited in Hossler et al., 1989). By 1998, Hossler, Schmidt, and Vesper found that students now start the search process in late spring or summer between their junior and senior years. It is also interesting to note that there are increasing numbers of ninth and tenth graders contacting colleges requesting information about the campus and major areas of study (Hossler et al., 1999).

By the year 2007, the number of high school graduates is projected to increase nationwide (Almanac, 1997). As the number of high school graduates increase, these graduates will bring new and distinct challenges for post-secondary institutions (Sevier, 1992). In general, this student population will be more urban, of lower SES status, and comprised of a larger minority population. As a result, college participation rates are likely to fall. All of these changes suggest that the prospective college population has changed dramatically in recent years. Thus, predicting application behavior based on models developed on historical data may be suspect. Therefore we need to refine our models that were built on earlier cohorts since these models were fitted on different student populations.

Methodology

Analytic Framework

To assist in the development of new information about student choice we compare the effectiveness of two analytic techniques in predicting student application behavior: artificial neural networks and logistic regression analysis. ANN research has primarily been used by engineers, statisticians, mathematicians, and cognitive psychologists, but has been virtually unexplored as a tool to aid educational researchers.

The structure of ANNs are based on how the human brain functions. The cells in the human brain provide humans with the ability to think, reason, and apply prior experiences to human action. The capability of the brain to learn is a function of the quantity of neurons in the brain and the various connections between the neurons. Within the human brain neurons have four basic components: dendrites, soma, axon, and synapses. Within a biological neuron dendrites receive inputs from the external environment, the soma processes the inputs received, and an axon converts the processed inputs into an output. And then the synapse creates an electrochemical contact from neuron to neuron. If the electrical charges sent by the receptors achieve a particular threshold level the nucleus of the neurons sends signals to other areas (e.g., muscles) within the human body. The exact threshold levels needed to trigger each neuron may be pre-wired at birth or learned throughout the years. It is this biological process of learning that ANN models mimic.

[FIGURE 1 ABOUT HERE]

Artificial neurons (see Figure 2) simulate the four basic components of the biological neuron, however, artificial neurons are far more simplistic than biological neurons. Similar to the human brain, an ANN learns to solve problems through experience. Within the ANN architecture, neurons (nodes) are divided into three layers: an input layer, a hidden layer, and an output layer. In traditional statistical nomenclature, the input layer is akin to the independent variables we include in regressions and the output layer is the outcome or dependent variable. The hidden layers consist of a variety of neurons that are connected to neighboring neurons by weights (see Figure 3). These weights act like coefficients in regression models. (In fact, Kuan and White (1994) demonstrated that linear, logistic, and probit regression models are special cases of ANNs.) Within an artificial neuron the inputs are processed (summed) within the cell body. If the summation exceeds the threshold value, the neuron sends an impulse to the axon (weight matrices).

[FIGURE 2 ABOUT HERE]

The weights are initially set to a small random number and as data is fed to the model the weights are adjusted using a feedback method (see Figure 3). The most common feedback method is known as back propagation (Rumelhart, Hinton, & Williams, 1986). The difference between the predicted and actual values (error) is fed back (back propagated) into the network and this process continues one layer at a time until the error is minimized to a prespecified level or the model is stopped by the researcher. This process permits the detection of trends and patterns in the data that would normally go undiscovered. Once trained the network can be supplied new or untrained data and classifications or predictions may be made. More specifically, the type of ANN classification utilized in this study is the multilayer perceptron (MLP). An MLP contains estimated weights between the inputs and the hidden layer and the hidden layer applies a nonlinear activation function. Multilayer perceptrons (MLPs) are layered feedforward networks and are commonly trained via backpropagation. (For a relatively non-technical treatment of ANN theory and application see Garson, 1998).

[FIGURE 3 ABOUT HERE]

<u>The Samples</u>

Data for this study was obtained from the admissions office of a large Research I institution located in Iowa. Two cohorts of information were used, each consisting of approximately 20,000 records. The cohorts were defined as all students who sent ACT scores to the study institution who were interested in applying for admission for the fall of 1998 and fall of 1999 entering classes. We used score senders as the potential pool of college-going students rather than high school graduates because we feel the former is more reflective of the actual pool of college prospects.

The data for this study was obtained from the admission's database that contains a wealth of information including student responses to the Student Profile Section of the ACT Assessment. The independent variables (listed in Table 1) were chosen based on prior research of student application behavior and knowledge about institution-specific factors. The dependent or output measure used in this study is an indicator of whether a student applied (1) or did not apply (0) to the study institution. The main objective was to test the ANN and logistic regression models predictive accuracy and use the results to provide information about student application propensities to institutional recruiters.

[TABLE 1 ABOUT HERE]

Results of the ANN and Logistic Models

The general analytic strategy used was to estimate (train) the logistic (ANN) model using the 1999 data set, and then test the predictive accuracy of each of these models on the 1998 data. In traditional statistical terms, the 1999 data is the "developmental" sample and the 1998 data is the validation (or sometimes known as the "holdout") sample. In ANN these are called training and testing data sets, respectively. We use this cross validation strategy because the classification or predictive accuracy of a model fit to the developmental (training) data set will tend to over predict how accurate the model is on other samples (Hosmer and Lemeshow, 1989).

The ANN models were estimated using SPSS's Clementine software, version 5.21 and the logistic regression model was fitted using SPPS version 9. The logistic model was estimated using maximum likelihood techniques, while the ANN models were estimated using the back propagation method.

Below we present the results of the competing approaches discussed above. The independent variables defined in Table 1 were used to fit the logistic regression model. Because ANNs are designed to uncover nonlinear patterns and are well-suited to dealing with missing data, some of the variables that were included as dummy specifications in the logistic regression model (e.g., ACT Composite score, High School Rank Percentile) were included as continuous variables in one of the ANN models trained. We also trained an ANN model using the exact same variables used in the logistic regression. We did so that we could compare how the models perform when each analytic technique was fed the same information.

Because our objective was to assist recruitment activities at the study institution we were primarily interested in how accurate the models predicted application to the institution. Even though it is not our primary interest, we also briefly discuss the relative importance of the inputs (independent variables) included in these models.

The results of the ANN model trained using the continuous input variables indicate that the model estimated on the developmental sample (1999 data) correctly classified 80.2 percent of prospective students (see Table 2). The results also indicate that 66.2 percent of applicants were correctly classified, and 88.4 percent of non-applicants were classified as such. As noted above, however, the developmental model overall correct classification rate (CCR) is overly optimistic when used to predict out-of-sample. When the trained ANN model was used to predicted application behavior on the holdout sample (1998 data), the CCR dropped to 77.8 percent. Relative to the training data set results, there was an improvement in predicting applicants in the test data set (66.2 vs. 70.9 respectively), but the prediction of non-applicants was lower in the holdout sample than in the developmental model (82.9 vs. 88.4, respectively).

[TABLE 2 ABOUT HERE]

The relative importance of the variables included in the continuous variable ANN model are presented in Table 3. The relative importance is similar to the partial correlation coefficients (denoted R in SPSS output) in logistic regression analysis. We find that high schools with historically high application to ACT score sender ratios have the highest level of importance in the ANN, followed by the variable measuring where the institution was in the students choice set, and the ACT score of the prospective student. As you will see, these constructs are also found to be highly related to application behavior in the other models estimated.

[TABLE 3 ABOUT HERE]

The results of the ANN model trained using the same variables used in the logistic regression model indicate that the model trained on the developmental sample (1999 data) had an overall correct classification rate (CCR) of 78 percent (see Table 4). The results indicate that 59.3 percent of applicants were correctly classified, and 89 percent of non-applicants were classified as such. Again, the developmental model overall CCR is an optimistic assessment of how well this model will predict out-of-sample. When this ANN model was used to predicted application behavior on the holdout sample (1998 data), the CCR dropped by two points to 76 percent. There was a slight improvement in predicting applicants, but the prediction of non-applicants was lower in the holdout sample than in the developmental model (87.2 vs. 89.0 percent, respectively).

[TABLE 4 ABOUT HERE]

The relative importance of the variables included in the model are presented in Table 5. The pattern is similar to that found when we used continuous variables to train the ANN, implying some consistency in the factors found to be related to student application behavior.

[TABLE 5 ABOUT HERE]

The classification results of the logistic model are presented in Table 6. The model built on the developmental sample (1999 data) had an overall correct classification rate (CCR) of 73.8 percent. The results indicate that 61.1 percent of applicants were correctly classified, and 85.5 percent of non-applicants were classified as such. The results of this model were validated on the 1998 data and compared to the developmental sample results the CCR dropped only slightly to 72.3 percent.

[TABLE 6 ABOUT HERE]

Table 7 presents the estimates of the independent variables used in the logistic regression model. A close inspection of the Wald or R (partial correlation) statistics reveals that most of the constructs identified in the ANN models as important are also the factors that exhibit powerful effects in the logistic regression model. Thus, we observe a great deal of consistency among the results of these different models.

In general, we found that the ANN trained using continuous variables had the highest CCR (78 percent), followed by the dummy variable ANN (76 percent) and then the logistic regression model (72 percent). Thus, for the samples used in this study, the logistic regression model does not predict application behavior as well as the ANN models do.

[TABLE 7 ABOUT HERE]

Implications and Conclusions

ANNs have several advantages over traditional statistical methods. First, ANNs do not require knowledge of the functional relationship between the independent variables and the dependent (i.e., the correct functional form or the degree of non-linearity) in order to estimate the model. Unlike traditional statistical techniques, ANNs learn from examples with a small number of prior assumptions about structural relationships. Second, ANNs are useful when logistic models are unable to fully recognize the complexities of the data. Third, ANNs are very good at pattern recognition and may accommodate related variables without incurring problems of multicollinearity (Etheridge, Sriram, & Hsu, 2000). Fourth, ANNs also perform well when there are massive amounts of data or the data are missing (Moore, 1988; Garson, 1998). Since ANNs treat input values are categories, under circumstances with noisy or missing data the ANN will determine the relative importance of each of the input values whether the values are legitimate or missing. Fifth, unlike logistic regression, ANNs detect nonlinearities and interactions automatically and can estimate multiple outputs at a time (Lee et al., 1993).

This is not to say that ANNs should replace traditional statistical methods, ANNs do have some disadvantages. First, the technique is a somewhat "black box" method in that the underlying theory and concepts are evolving and are not well understood by many researchers. Second, ANNs can require considerable training time. One variant of the continuous variable ANN ran for many hours. What the researcher must do is to become adept at when to manually stop the models (unlike regression you can choose to stop the training and still obtain results). Sometimes the extra time spent letting the model train is not worth the minimal increase in predictive accuracy. Third, ANNs make it difficult to determine the relationship between the dependent and independent measure. For instance, ANNs may provide information that two independent variables are important, however, the ANN will not distinguish how applicants and non-applicants are distinct with regard to these independent variables. Logistic regression models, on the other hand, are very good at distinguishing among the various factors and characteristics of applicants and non applicants, and typically do so very quickly. Finally, the implementation of an ANN can be difficult process (Hecht-Nielson, 1990). Considerable time was spent in learning the software, although it appears that recent improvements in this package and others may make that less of a concern in the immediate future.

With regard to institutional policy, the results of this analytic effort are being used at the study institution to assist institutional recruitment efforts. The continuous ANN model was used to predict the application behavior of students who sent ACT scores to the study institution and are considering college enrollment in the fall of 2002. Estimating a probability of application for each prospective applicant allows the admission's staff to prioritize and target their recruitment efforts on particular groups of students. Doing so allows recruiters to determine whether to scale down mailings and/or telephone contacts on groups who are poor application prospects. This targeted approach has the potential of saving considerable resources given that more than 20,000 students typically send scores to the study institution in any year. Also, the telemarketing staff has used the predictions to group prospective students into deciles and prioritize telephone campaigns. This is a far cry from the rather ad hoc approach used in the past.

We also used the predictions to reevaluate how the institution buys names from vendors like Educational Testing Service. Using the predictions with other variables in the admission's database we were able to eliminate some students from tape buys, resulting in a moderate savings of admissions recruitment funds.

One other benefit of this project was that while we were learning about ANNs and sharing our successes and frustrations with colleagues and administrators, we learned that some engineering colleagues are using artificial intelligence (AI) to study other outcomes. Not only is there a cadre of AI researchers right next door to us, but they have also established and AI lab to do research and train individuals in the use of these new techniques. With the help of these experts we are beginning to learn more about the technique and are beginning to train College of Education students in the use of these methods. Also, one of the authors of this paper and an engineering faculty member wrote an NSF grant to help fund the training of colleagues and students.

While ANNs and logistic regression each have their advantages and disadvantages, a combination of the two techniques provides institutional researchers with new ways to further refine predictive accuracy, generalization, and model fit (Hung, Hu, Patuwo, and Shanker, 1996; Richard & Lippman, 1991). Issues regarding application behavior must be critically and analytically mapped out as a measure of ensuring optimal resource allocation and development. These efforts can help to make educational planning more efficient and effective. However, as mentioned earlier, interest in neural networks has remained unexplored by educational researchers and policy analysts. Hopefully our venture into this new area of modeling will encourage other institutional researchers to use this technique to explore the complex processes found in our educational institutions.

REFERENCES

Almanac. (1997). The Chronicle of Higher Education, 44, 6-7, 19.

Bateman, M., & Spruill, D. (1996). Student decision-making: Insights from the college choice process. College Student Journal, 30, 182-186.

Bruggink, T. H., & 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).

Bryant, P., & Crockett, K. (1993). The admissions office goes scientific. Planning for Higher Education, 22, 1-8.

Chapman, D. W. (1981). A model of student college choice. Journal of Higher Education, 52(5), 490-505.

Cybenko, G. (1989). Approximation by superpositions of a sigmoid function. Mathematical Control Signals Systems, 2, 303-314.

Dehne, G. C. (1994). Reinventing student recruitment. Trusteeship, 2(3), 11-15.

--------------- (1999). Modeling the college application decision process in a land-grant university. Economics of Education Review, 18, 117-132.

Dey, E. L., Astin, A. W., & Korn, W. S. (1991). The American freshman: twenty-five year trends, 1966-1990. Los Angeles: Higher Education Research Institute, UCLA.

Etheridge, H. L., Sriram, R. S., & Hsu, H.Y. K. (Spring 2000). A comparison of selected artificial neural networks that help auditors evaluate client financial viability. Decision Sciences, 31 (2), 531-549.

Freeman, H. B. (1984). Impact of no-need scholarships on the matriculating decision of academically talented students. Paper presented at the annual meeting of the American Association of Higher Education, Chicago.

Galotti, K. M., & Mark, M. C. (1994). How do high school students structure an important life decision? A short-term longitudinal study of the college decision-making process. Research in Higher Education, 17, 589-607.

Ganderton, P. T. (1992). The effect of subsidies in kind on the choice of college. The Journal of Public Economics, 48, 269-292.

Garson, G. D. (1998). Neural networks: An introductory guide for social scientists. Thousand Oaks: Sage Publications Inc.

Goyette, K. (1999). Application to college: A comparison of Asian American and White high school students. Paper presented at the annual meeting of the American Educational Research Association, Montreal. ED432 623.

Hamrick, F., & Hossler, H. (1996). Diverse information-gathering methods in the post secondary decision-making process. Review of Higher Education, 19, 179-198.

Hearn, J. (1984). The relative roles of academic ascribed and socioeconomic characteristics in college destinations. Sociology of Education, 57, 22-30.

Hecht-Nielsen, R. (1990). Neurocomputing. Reading: Addision-Wesley.

Hosmer, D. W. and Lemeshow, S. (1989). Applied regression analysis. New York: John Wiley and Sons.

Hossler, D., & Gallagher, K. S. (1987). Studying student college choice: A three-phase model and the implications for policymakers. College and University, 2, 207-221.

Hossler, D., Braxton, J., & Coopersmith, G. (1989). Understanding student college choice. In J. C. Smart (Ed.), Higher education: Handbook of theory and research Vol. 5 (pp. 231-288). New York: Agathon Press.

Hossler, D., Schmit, J., & Vesper, N. (1999). Going to college: How social, economic, and educational factors influence the decisions students make. Baltimore: The John Hopkins University Press.

Hung, M. S., Hu, M. Y., Patuwo, B. E., & Shanker, M. (1996). Estimating posterior probabilities in classification problems with neural networks. International Journal of Computational Intelligence and Organizations, 1, 49-60.

Ihlanfeldt, W. (1980). Achieving optimal enrollments and tuition revenues. San Francisco: Jossey-Bass.

Kitano, H. H.L., & Daniels, R. (1988). Asian Americans: Emerging minorities. Englewood Cliffs: Prentice Hall.

Kohn, M. G., Manski, C. F., & Mundel, D. S. (1976). An empirical investigation of factors which influence college going behavior. Annals of Economic and Social Measurement, 5(4), 391-418.

Kolstad, A. J. (1979). The influence of high school type and curriculum on enrollment in higher education and postsecondary training. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ED 173627)

Kuan, C. and White, H. (1994). Artificial neural networks: An econometric perspective. Econometric Reviews, 13(1): 1-91.

Lee, T., White, H., and Granger, Clive W.J. (1993). Testing for neglected nonlinearity in time series models: a comparison of neural network methods and alternative tests. Journal of Econometrics, 56 (3), 269-91.

Litten, L. H., Sullivan, D., & Brodigan, D. L. (1983). Applying market research in college admissions. New York: College Entrance Examination Board.

Litten, L. H. (1982). Different strokes in the applicant pool: Some refinements in a model of student college choice. Journal of Higher Education, 53(4), 383-402.

Manski, C. F., & Wise, D. A. (1983). College choice in America. Cambridge: Harvard University Press.

McDonough, P. M. (1994). Buying and selling in higher education: The social construction of the college applicant. Journal of Higher Education, 65(4), 427-446.

McLanahan, S., & Sandefur, G. (1994). Growing up with a single parent: What hurts, what helps. Cambridge: Harvard University Press.

Melia, P., & Goodman, M. (1988). Controlling demographic impacts. College and University, 64, 51-63.

Min, P. (1988). The Korean Family. In C. H. Mindel, R. W. Habenstein, & R. Wright, Jr. (Eds.), Ethnic families in America: Patterns and variations (pp. 199-299). New York: Elsevier.

Paulsen, M. B. (1990). College choice: Understanding student enrollment behavior. ERIC Digest, 6. Washington, DC: ASHE-ERIC Higher Education Reports.

Patuwo, B. E., Hu, M. Y., & Hung, M. S. (1993). Two group classification using neural networks. Decision Sciences, 24(4), 825-845.

Rainsford, G. N. (1985). Enrollment management: A president's perspective. College and University, 60, 336-344.

Richard, M. D., & Lippman, R. P. (1991). Neural network classifiers estimate Baysesian a posterior probabilities. Neural Computation, 3, 461-483.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, & J. L. Williams (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition. Cambridge: MIT Press.

Sanoff, A. P. (1994). Admissions deans on hot seat. U.S. News and World Report, 117, 98-101.

Sevier, R. A. (1992). Is demography destiny? Journal of College Admissions, 135, 13-22.

Shea, C. (1994). A flood of applications. Chronicle of Higher Education, 40, A31-A32.

Stanton-Salazar, R. D., & Dornbusch, S. M. (1995). Social capital in the reproduction of inequality: Information networks among Mexican-origin high school students. Sociology of Education, 68, 116-35.

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

Welki, A. M., & Navratil, F. J. (1987). The role of applicants' perceptions in the choice of college. College and University, Winter, 147-160.

Figure 1: Four Basic Components of a Human Biological Neuron

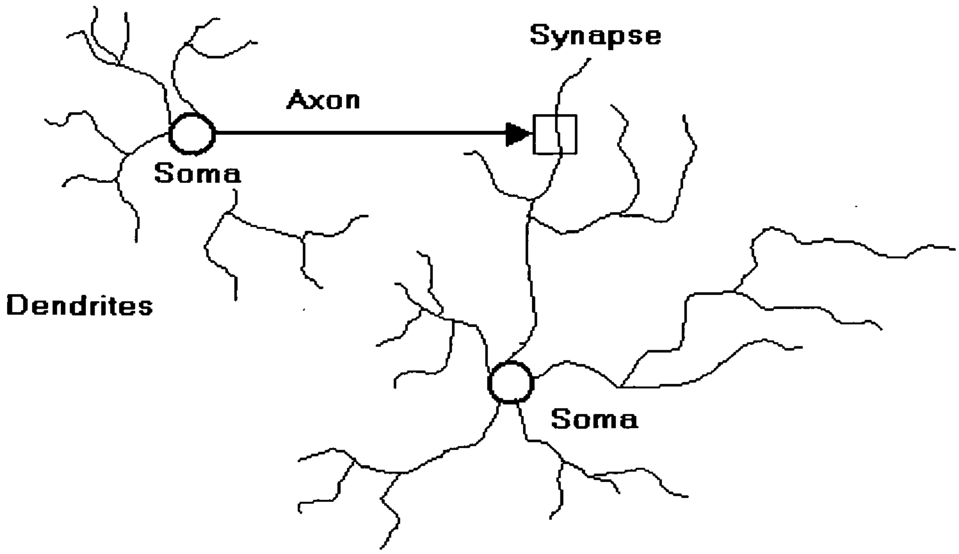

Figure 2: The Components of a Basic Artificial Neuron

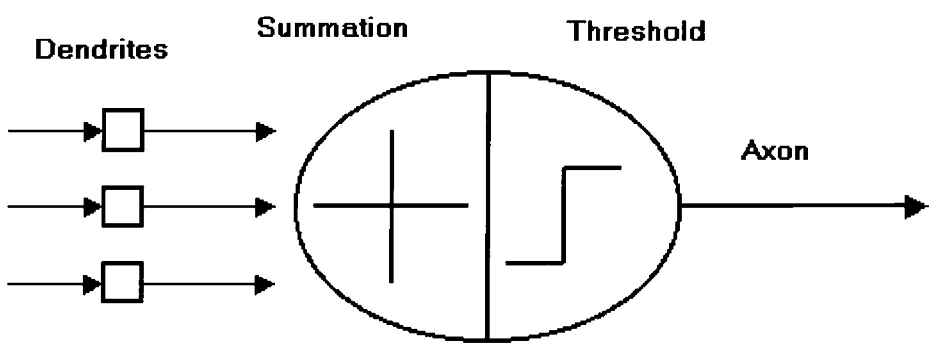

Figure 3: Chaining of Operations in Back Propagation

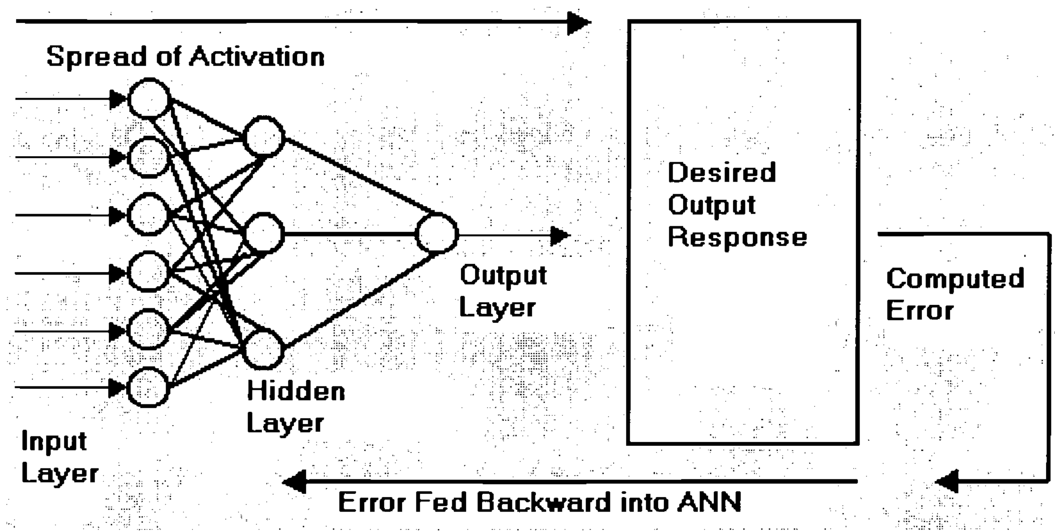

Table 1: Definition of the Explanatory Variables

Variable Name	Definition
Demographic Variables	
Gender	
Female	A dummy equal to one if the student is female
Race/Ethnicity	
Native American Indian	A dummy (=1) if Native American/Alaskan Native
African American	A dummy equal to one if the student is African American
Hispanic	A dummy equal to one if the student is Hispanic
Asian American	A dummy (=1) if the student is Asian or Pacific Islander
Multiracial	A dummy equal to one if the student is multiracial
Caucasian	**Reference Group.** A dummy equal to one if the student is Caucasian, had missing data, or preferred not to respond
Family Income	
Top Income Quartile	A dummy (=1) if family income is \$80,000-\$100,000
Second Income Quartile	A dummy (=1) if family income is \$60,000-\$80,000
Third Income Quartile	A dummy (=1) if family income is \$42,000-\$60,000
Missing Income	A dummy equal to one if family income is missing
Bottom Income Quartile	**Reference group.** A dummy (=1) if income < \$42,000
Community Size	
Farm	**Reference group.** A dummy (=1) if from a farm
Less than 500	A dummy (=1) if community size is less than 500 people
500-1,999	A dummy (=1) if community size 500-1,999
2,000-9,999	A dummy (=1) if community size 2,000-9,999
10,000-49,999	A dummy (=1) if community size 10,000-49,999
50,000-249,999	A dummy (=1) if community size 50,000-249,999
250,000-499,999	A dummy (=1) if community size 250,000-499,999
500,000-999,999	A dummy (=1) if community size 500,000-999,999
One Million or more	A dummy (=1) if community size one million or more
Missing Community	A dummy (=1) if the student's community size is missing
High School Variables	
ACT Composite Quartiles	
Bottom ACT Quartile	**Reference Group.** A dummy (=1) if student's ACT composite is 0-19
Third ACT Quartile	A dummy (=1) if student's ACT composite is 20-23
Second ACT Quartile	A dummy (=1) if student's ACT composite 24-26
Top ACT Quartile	A dummy (=1) if student's ACT composite 27-36
High School GPA	
GPA	A dummy (=1) if the student's HS GPA was A or B
HS Course Work	
English	A dummy (=1) if the student had four years of English
Math	A dummy (=1) if the student had three years of Math
Social Studies	A dummy (=1) if the student had 2 years of Social Studies
Natural Sciences	A dummy (=1) if the student had 3 yrs of Natural Sciences

Spanish	A dummy (=1) if the student had two years of Spanish
German	A dummy (=1) if the student had two years of German
French	A dummy (=1) if the student had two years of French
High School Type	
Catholic High School	A dummy (=1) if the student attended a Catholic HS
Graduating Class Size	
Fewer than 25	**Reference Group.** If HS class contained < 25 students
25-99	A dummy (=1) if HS graduating class size 25-99 students
100-199	A dummy (=1) if HS graduating class size 100-199
200-399	A dummy (=1) if HS graduating class size 200-399
400-599	A dummy (=1) if HS graduating class size 400-599
600-899	A dummy (=1) if HS graduating class size 600-899
900 or more	A dummy (=1) if HS graduating class size 900+ students
High Yield Applicants	
High Yield High School	A dummy (=1) if student is from a HS with a historical ratio of applicants to test score senders >= .75
College Preference Variables	
Distance from Institution	
Less than 10 miles	**Reference Group.** A dummy (=1) if student's home residence is less than ten miles from the institution
10-25 miles	A dummy (=1) if distance is 10 to 15miles
26-100 miles	A dummy (=1) if distance is 26 to 100 miles
More than 100 miles	A dummy (=1) if distance is more than 100 miles
Other	A dummy (=1) if distance missing/not decided on college
Tuition (Does not include room & board)	
$500	A dummy (=1) if expects yearly tuition of $500.
$1,000	A dummy (=1) if expects yearly tuition of $1,000
$2,000	A dummy (=1) if expects yearly tuition of $2,000
$3,000	**Reference Group.** If expects yearly tuition of $3,000
$4,000	A dummy (=1) if expects yearly tuition of $4,000
$5,000	A dummy (=1) if expects yearly tuition of $5,000
$7,000	A dummy (=1) if expects yearly tuition of $7,000
$10,000	A dummy (=1) if expects yearly tuition of $10,000
Missing/Other	A dummy (=1) missing data/ no tuition preference
University College Choice	
First Choice	A dummy equal to one if institution is first choice college
Second Choice	A dummy equal to one if institution is second choice
Third Choice	A dummy equal to one if institution is third choice
Supplemental	A dummy equal to one if student sends supplemental scores
Public Institution	A dummy (=1) if student plans to attend public institution
Major	
Agriculture Sciences	A dummy (=1) if planned major in Agriculture Sciences
Business and Marketing	A dummy (=1) if planned major in Business or Marketing
Communications	A dummy (=1) if planned major in Communications
Architecture	A dummy (=1) if planned major in Architecture

Community Related	A dummy (=1) if planned major in Community/Personal Services; Human, Family, Consumer Service or a Trade
Education	A dummy (=1) if planned major in Education
Math/Comp. Science	A dummy (=1) if planned major in Math/Computer Science
Liberal Arts	A dummy (=1) if planned major in Cross Disciplinary Studies; Foreign Language; Letters; Philosophy, Religion, and Theology; Visual and Performing Arts
Engineering	A dummy (=1) if planned major in Engineering
Sciences	A dummy (=1) if planned major in Health Sciences, Allied Health, Bio or Physical Sciences, Social Sciences
Undecided/Missing Major	A dummy (=1) if the student is undecided/did not respond
Varsity Athletics	A dummy (=1) if interest in participating in varsity athletics
Fraternity/Sorority	A dummy (=1) if interest in fraternity or sorority
College Size	A dummy (=1) if interest in college of 20,000+ students
Highest Expected Education Level	
Vocational	A dummy (=1) if interested in vocational degree
Two Year Degree	A dummy (=1) if interested in obtaining a two year degree
Bachelor Degree	**Reference Group.** A dummy equal to one if the student is interested in obtaining a bachelor's degree
Graduate Degree	A dummy (=1) if interested in obtaining a graduate degree
Professional Degree	A dummy (=1) if interest in obtaining a professional degree
Missing Expectations	A dummy (=1) if educational expectation is missing
Public Institution	A dummy (=1) if interested in attending a public institution

Table 2: Classification Effectiveness of the Continuous Variable Artificial Neural Network Model

Classification Table for Model Trained on 1999 Data

		Predicted			
		Applied		Did Not	
		N	%	N	%
Actual	Applied	4853	**66.2**	1451	11.6
	Did Not	2479	33.8	11074	**88.4**

Overall Correct Classification Rate 80.2

Accuracy of Classification Using Trained Model on the Holdout Sample

		Predicted			
		Applied		Did Not	
		N	%	N	%
Actual	Applied	5280	**70.9**	1766	17.1
	Did Not	2171	29.1	8531	**82.9**

Overall Correct Classification Rate 77.8

Table 3: Continuous Variable Artificial Neural Network Model Results

Overall Predicted Accuracy	80.2%
Structure of the Model	
Input Layer	170 neurons
Hidden Layer #1	20 neurons
Hidden Layer #2	15 neurons
Hidden Layer #3	10 neurons
Output Layer	1 neuron
Relative Importance of the Inputs	
High Yield High School	0.674
Choice	0.500
ACT Score	0.255
Years of HS French	0.155
Distance from Institution	0.146
Community Size	0.137
Race/Ethnicity	0.122
Preferred Tuition Level	0.119
Years of HS Math	0.119
Years of HS Spanish	0.118
Years of HS German	0.118
Major	0.116
Years of HS Natural Science	0.115
Years of HS English	0.115
Family Income	0.113
Interested in Frat/Sorority	0.095
High School Type Attended	0.091
Preferred College Size	0.080
Preferred College Type	0.077
Years of HS Social Science	0.072
Gender	0.071
Expected Education Level	0.056
Interested in Varsity Athletics	0.054
High School GPA	0.030
Size of HS Graduating Class	0.017

Table 4: Classification Effectiveness of the Dummy Variable Artificial Neural Network Model

Classification Table for Model Trained on 1999 Data

		Predicted: Applied N	Predicted: Applied %	Predicted: Did Not N	Predicted: Did Not %
Actual	Applied	4344	**59.3**	1384	11.0
	Did Not	2988	40.7	11141	**89.0**

Overall Correct Classification Rate 78.0

Accuracy of Classification Using Trained Model on the Holdout Sample

		Predicted: Applied N	Predicted: Applied %	Predicted: Did Not N	Predicted: Did Not %
Actual	Applied	4522	**60.7**	1323	12.8
	Did Not	2929	39.3	8974	**87.2**

Overall Correct Classification Rate 76.0

Table 5: Dummy Variable Artificial Neural Network Model Results

Overall Predicted Accuracy	78.0%
Structure of the Model	
Input Layer	75 neurons
Hidden Layer #1	20 neurons
Hidden Layer #2	15 neurons
Hidden Layer #3	10 neurons
Output Layer	1 neuron
Relative Importance of the Inputs	
Supplemental Choice	0.5958
High Yield HS	0.4223
First Choice	0.2784
2nd Qrt. ACT	0.1188
Top Qtr. ACT	0.1109
Communications	0.0899
Top Qtr. Income	0.0808
Multiracial	0.0789
Latino/a	0.0764
Comm. Size to 500	0.0718
Community, Human, Personal Serv.	0.0650
3rd Qrt. ACT	0.0641
Distance 10 to 25	0.0627
Second Choice	0.0572
AA Degree	0.0525
HS Grades A or B	0.0522
Agricultural Science/Tech	0.0504
HS Size Missing	0.0485
Tuition to $1000	0.0483
Architecture/Environmental Design	0.0459
Engineering	0.0454
Prefer Public Institution	0.0454
2 Yrs French	0.0453
Attended Catholic HS	0.0440
Graduate Degree	0.0432
Comm. Size to 249999	0.0420
Comm. Size Missing	0.0392
Tuition Missing	0.0380
Income Missing	0.0374
Education	0.0367

Table 5: Dummy Variable Artificial Neural Network Model Results (Cont'd)

Relative Importance of the Inputs	
Professional Degree	0.0364
HS Size to 100	0.0352
Tuition to $5000	0.0338
Distance Missing	0.0304
American Indian	0.0304
Interest Frat/Sor.	0.0299
Math/Computer Science	0.0286
Prefer College GT 20K	0.0282
2 Yrs. German	0.0268
Comm. Size to 49999	0.0267
HS Size to 200	0.0266
Distance GT 100	0.0262
Tuition to $7500	0.0260
HS Size to 25	0.0258
4 Yrs. HS English	0.0256
Interest Varsity Ath.	0.0251
Gender	0.0234
Gender	0.0234
Tuition to $10000	0.0231
Health, Sciences, Social Science	0.0228
Comm. Size to 1999	0.0220
2 Yrs. Social Science	0.0193
Asian American	0.0193
Distance 26 to 100	0.0191
HS Size to 400	0.0185
Third Choice	0.0184
Business/Marketing	0.0177
Tuition to $500	0.0174
Tuition to $4000	0.0172
HS Size to 900	0.0163
Liberal Arts	0.0154
Comm. Size to 499999	0.0151
Expected Ed. Missing	0.0150
Comm. Size 1M Plus	0.0145
3 Yrs. HS Math	0.0143
2 Yrs. Spanish	0.0126
2nd Qrt. Income	0.0124
Undecided Major	0.0119

Table 5: Dummy Variable Artificial Neural Network Model Results (Cont'd)

Relative Importance of the Inputs	
Comm. Size to 9999	0.0116
HS Size to 600	0.0114
Comm. Size to 999999	0.0110
3 Yrs. Natural Science	0.0102
African American	0.0100
3rd Qrt. Income	0.0090
Voc Tech Degree	0.0069
Tuition to $2000	0.0055

Table 6: Classification Effectiveness of the Logistic Regression Model

Classification Table for Logistic Model Using the 1999 Data

		Predicted: Applied N	Predicted: Applied %	Predicted: Did Not N	Predicted: Did Not %
Actual	Applied	3653	**61.1%**	945	14.5%
	Did Not	2325	38.9%	5551	**85.5%**

Overall Correct Classification Rate 73.8

Accuracy of Classification Using Fitted Model on Holdout Sample

		Predicted: Applied N	Predicted: Applied %	Predicted: Did Not N	Predicted: Did Not %
Actual	Applied	3734	**63.3%**	922	17.5%
	Did Not	2167	36.7%	4346	**82.5%**

Overall Correct Classification Rate 72.3

Table 7: Logistic Regression Model Results

	Chi-square	df	Significance
Model	4802.67	74	0.000

Variable	B	S.E.	Wald	Significance	R	Odds Ratio
Gender	0.192	0.052	13.758	0.000	0.027	1.211
American Indian	-0.117	0.466	0.063	0.802	0.000	0.890
African American	0.318	0.150	4.506	0.034	0.012	1.374
Latino/a	0.575	0.182	9.948	0.002	0.022	1.776
Asian American	0.418	0.153	7.480	0.006	0.018	1.519
Multiracial	-4.866	3.049	2.546	0.111	-0.006	0.008
Top Qtr. Income	0.470	0.085	30.214	0.000	0.042	1.599
2nd Qrt. Income	0.202	0.080	6.486	0.011	0.017	1.224
3rd Qrt. Income	0.020	0.082	0.061	0.805	0.000	1.020
Income Missing	0.259	0.106	5.971	0.015	0.016	1.295
Comm. Size to 500	-0.177	0.165	1.152	0.283	0.000	0.838
Comm. Size to 1999	0.069	0.110	0.393	0.531	0.000	1.071
Comm. Size to 9999	0.063	0.098	0.418	0.518	0.000	1.065
Comm. Size to 49999	0.169	0.101	2.769	0.096	0.007	1.184
Comm. Size to 249999	0.292	0.104	7.865	0.005	0.019	1.339
Comm. Size to 499999	0.172	0.142	1.476	0.224	0.000	1.188
Comm. Size to 999999	-0.056	0.221	0.064	0.800	0.000	0.946
Comm. Size 1M Plus	0.392	0.182	4.655	0.031	0.013	1.479
Comm. Size Missing	0.637	0.175	13.205	0.000	0.026	1.892
Top Qtr. ACT	0.862	0.085	103.224	0.000	0.079	2.368
2nd Qrt. ACT	0.920	0.077	144.451	0.000	0.093	2.509
3rd Qrt. ACT	0.640	0.070	83.748	0.000	0.071	1.896
HS Grades A or B	0.523	0.066	63.525	0.000	0.061	1.687
4 Yrs. HS English	0.208	0.078	7.110	0.008	0.018	1.232
3 Yrs. HS Math	0.036	0.065	0.318	0.573	0.000	1.037
2 Yrs. Social Science	-0.144	0.129	1.247	0.264	0.000	0.866
3 Yrs. Natural Science	0.034	0.054	0.402	0.526	0.000	1.035
2 Yrs. Spanish	-0.046	0.058	0.642	0.423	0.000	0.955
2 Yrs. German	0.132	0.159	0.696	0.404	0.000	1.142
2 Yrs French	-0.337	0.135	6.233	0.013	-0.016	0.714
HS Size to 25	0.735	0.193	14.475	0.000	0.028	2.086
HS Size to 100	0.633	0.196	10.484	0.001	0.023	1.884
HS Size to 200	0.820	0.194	17.937	0.000	0.031	2.270
HS Size to 400	0.887	0.198	20.183	0.000	0.033	2.428
HS Size to 600	0.945	0.210	20.185	0.000	0.033	2.574
HS Size to 900	1.095	0.227	23.341	0.000	0.036	2.988
HS Size Missing	2.125	0.298	50.915	0.000	0.055	8.371

Table 7: Logistic Regression Model Results (Cont'd)

Variable	B	S.E.	Wald	Significance	R	Odds Ratio
High Yield HS	1.983	0.149	178.222	0.000	0.104	7.262
Distance 10 to 25	-0.369	0.172	4.625	0.032	-0.013	0.691
Distance 26 to 100	0.211	0.116	3.320	0.068	0.009	1.235
Distance GT 100	0.483	0.109	19.592	0.000	0.033	1.621
Distance Missing	0.188	0.110	2.899	0.089	0.007	1.206
Tuition to $500	-0.269	0.540	0.247	0.619	0.000	0.764
Tuition to $1000	0.077	0.337	0.053	0.819	0.000	1.080
Tuition to $2000	-0.198	0.177	1.255	0.263	0.000	0.820
Tuition to $4000	-0.048	0.128	0.141	0.707	0.000	0.953
Tuition to $5000	0.017	0.109	0.023	0.879	0.000	1.017
Tuition to $7500	0.061	0.112	0.292	0.589	0.000	1.062
Tuition to $10000	0.201	0.115	3.073	0.080	0.008	1.223
Tuition Missing	0.005	0.088	0.004	0.951	0.000	1.005
First Choice	1.838	0.075	595.479	0.000	0.190	6.284
Second Choice	0.702	0.078	82.052	0.000	0.070	2.018
Third Choice	0.320	0.082	15.165	0.000	0.028	1.378
Supplemental Choice	4.141	0.135	935.724	0.000	0.239	62.861
Intcrest Varsity Ath.	-0.214	0.051	17.753	0.000	-0.031	0.808
Interest Frat/Sor.	0.164	0.049	11.417	0.001	0.024	1.178
Prefer College GT 20K	0.300	0.056	28.256	0.000	0.040	1.350
Voc Tech Degree	-0.186	0.615	0.091	0.762	0.000	0.830
AA Degree	-0.768	0.290	6.996	0.008	-0.017	0.464
Graduate Degree	0.271	0.062	19.261	0.000	0.032	1.311
Professional Degree	0.286	0.064	20.186	0.000	0.033	1.331
Expected Ed. Missing	-0.040	0.173	0.053	0.818	0.000	0.961
Prefer Public Institution	0.251	0.066	14.539	0.000	0.028	1.285
Undecided/Miss. Major	0.158	0.167	0.893	0.345	0.000	1.171
Ag Sciences/Tech.	-0.552	0.290	3.624	0.057	-0.010	0.576
Architecture	-0.344	0.239	2.081	0.149	-0.002	0.709
Business/Marketing	0.159	0.152	1.095	0.295	0.000	1.173
Communications	0.177	0.118	2.228	0.136	0.004	1.193
Comm/Personal Services	-0.280	0.162	2.995	0.084	-0.008	0.756
Math/Computer Science	-0.203	0.134	2.271	0.132	-0.004	0.817
Liberal Arts	0.163	0.177	0.847	0.358	0.000	1.177
Education	-0.132	0.174	0.574	0.449	0.000	0.876
Engineering	0.006	0.172	0.001	0.972	0.000	1.006
Health, Social Sciences	0.227	0.157	2.091	0.148	0.002	1.255
Constant	-5.121	0.304	284.115	0.000		

-2 Log Likelihood	11618.35
Cox & Snell R^2	0.32
Nagelkerke R^2	0.437

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").