Open Access Library Journal
2018, Volume 5, e4499
ISSN Online: 2333-9721
ISSN Print: 2333-9705

A Predictive Model for Graduate Application to Enrollment

Vahid Lotfi[1*], Bradley Maki[2]

[1]School of Management, University of Michigan Flint, Flint, Michigan, USA
[2]Office of Graduate Programs, University of Michigan Flint, Flint, Michigan, USA
Email: *vahid@umflint.edu, bmaki@umflint.edu

How to cite this paper: Lotfi, V. and Maki, B. (2018) A Predictive Model for Graduate Application to Enrollment. *Open Access Library Journal*, **5**: e4499.
https://doi.org/10.4236/oalib.1104499

Received: March 12, 2018
Accepted: April 17, 2018
Published: April 20, 2018

Abstract

This study involved an investigation of factors that affect a graduate applicant in accepting an offer of admission and enrolling in a graduate program of study at a mid-sized public university. A predictive model was developed, using Decision Tree methodology to assess the probability that an admitted student would enroll in the program during the semester following acceptance. The study included actual application information such as demographic information, distance from the campus, program of interest, tests scores, financial aid, and other pertinent application items of over 4600 graduate applications over a three-year period. The Decision Tree model was then compared with a Bayesian Network model to reaffirm its validity and its predictive power. The method with the more promising outcome was used to develop predictive models for applicants interested in a sample of academic majors. The results of the predictive models were used to illustrate development of recruitment strategies for all applicants as well as for those interested in specific majors.

Subject Areas

Big Data Search and Mining, Mathematical Analysis, Mathematical Statistics

Keywords

Graduate Education, Student Recruiting, Predictive Modeling, Enrollment Management

1. Introduction

Many colleges and universities in the United States are experiencing overall enrollment decline due to a decreasing number of high school graduates [1].

Some colleges and universities, particularly public institutions, are also experiencing decline in their financial resources in general and for recruiting and offers of financial aid in particular [2] [3] [4]. The combination of declining enrollment and reduced financial resources requires more advanced and data driven strategic enrollment planning and management. Effective enrollment planning strategies are also essential for optimal capacity planning with respect to instructional as well as student service resources.

The origins of strategic enrollment management date back to the 1970s when some college admissions officers developed strategies to maintain their enrollment levels to mitigate projected decreases in the number of high school graduates [5]. As the field of "managing" enrollment evolved over time, it involved analysis of demographic data, segmentation of student populations, and more focused marketing to prospective students.

An integral and key element of strategic enrollment management is the development of effective enrollment forecasts. Reference [4] presents an overview of methodologies used in enrollment forecasting. Predictive models, based on various statistical techniques, are often used for this purpose. Such models can be used not only in enrollment forecasting but also to develop recruitment and retention strategies. A common approach to building a predictive model is to target specific sub-populations of students to forecast enrollment. One such population is graduate students. Although a number of articles focusing on predicting undergraduate enrollment have been published in the literature, the authors have not found any significant source with an emphasis on graduate enrollment.

2. Related Literature

Reference [6] presented an argument for identifying students for whom recruitment efforts should have the most impact. They developed a model using logistic regression that predicted the probability of a full-time admitted freshman student would enroll based on four groups of variables: demographic, academic, geographic, and behavior. The model was developed based on three years of data and was applied in an experiment where the admissions office tested the efficacy of increased contact with students who had enrollment probabilities between 30% and 60% (the middle of the distribution). The results of the experiment found a statistically significant difference in the actual enrollment of the experimental group as compared to the control group. This study did not include any financial aid data, possibly limiting the potential predictive power of the model.

Some institutions use statistical techniques developed by consulting firms to estimate probability of enrollment [7]. However, "… there is scant literature available with regard to how this analysis actually takes place" ([7], p. 532) due to reluctance to divulge recruitment strategies that are effective. The author developed a predictive model, using logistic regression, for assessing the probability that an applicant would actually enroll. The study was conducted at a rela-

tively large research institution. The dataset was limited to admitted undergraduate students with early admission in first week of January, applicants who were not recruited athletes, and applicants for whom the institution had ACT Student Profile Questionnaire (SPQ) information. The model predictors included demographic information, high school characteristics, family income level, and application date. The data for the academic year 1999 was used to develop and validate the model using 50/50 split sampling and threshold probability of 0.5. Reference [7]'s study prompted the author to recommend that institutions should prioritize their recruiting efforts on students who are wavering on whether or not to enroll, though he acknowledges that the model alone cannot tell managers who these students are exactly and that they need to also rely on the expertise of enrollment management professionals.

A predictive model for determining the probability that a student who has inquired about undergraduate programs at their institution would actually enroll in the following fall term was developed by [8]. A sample of 15,827 inquiries that had been received in 2003 was used to develop the model. The authors used the Bayesian Model Averaging (BMA) technique to develop the model. A number of geodemographic variables and "contact" variables were used as predictors. For students who were missing some of the geodemographic variables, the student's zip code generic census characteristics were used. Split sampling with a 50/50 split ratio was used to assess the prediction power of the models. The most promising model resulted in an overall prediction power of 89.25% with a sensitivity of about 36% and specificity of 97% assuming positive response if the probability to enroll was 0.5 or more. This model, while promising, was focused exclusively on undergraduate inquiries.

Another study [9] presented a predictive model for determining enrollment at a small, private liberal arts college in Wisconsin. A limited small number of biodemographic, athletic, contact, and "choice" variable were studied using logistic regression. Lodesma found that academically strong applicants had more options on where to attend and also more likely to apply to other schools and attend them. He found that the college may be attracting students who are already settled on attending. Using the 50/50 split ratio to assess the prediction power, the model resulted in a 64% overall correct classification (prediction power) with a sensitivity of 21% and specificity of 43%. This model was important in showing the differences when a school has a specific mission and profile, rather than a large, public research university. However, the number of variables studied was small and it too was focused on undergraduate enrollment.

Reference [10] examined several data mining techniques for predicting enrollment of accepted international applicants at large metropolitan Australian University. Their dataset consisted of 24,283 offers of admission made to international students between academic years 2008 and 2013. A set of 29 attributes (predictors) was included in the dataset, categorized into geodemographic, academic, and application-related items. The authors first conducted a correlation analysis to assess the relationship between the acceptance status (*i.e.*, enrollment

status) representing the dependent variable and the predictors. They then employed Principal Component analysis to determine the dimensionality of dependent variable as well as the relative strength of each of the predictors. Ten components were identified with the first two accounting for 99.7% of the variance. Country of citizenship, number of days between application and start of the term, major, and school/college had the most significant weights on the first two components. The authors also used several data mining techniques and compared their predictive powers. Logistic regression and neural networks had the most predictive powers with approximately 67.6% and 68.1% overall correct classifications, respectively. Their study was focused on international applications only and did not include domestic students.

3. Problem Definition

We considered the problem of developing a predictive model for determining whether a graduate student, admitted to a program of study, would actually register during the semester for which he/she had been admitted. Our study institution was a regional campus of a large public research university, classified as a Master's Colleges & Universities: Larger Programs according to the Carnegie Classification of Institutions of Higher Education. The campus was located in an urban setting in the upper Midwest in a city with a population of approximately 97,000 people. At the time of this study, there were 38 distinct graduate programs available. The student body was composed primarily of people who lived within a commutable distance of the physical campus with smaller populations of online students and international students.

The dataset in our study was obtained from the graduate admissions office. The university enrolled between 8000 and 8500 students overall in 2015 and 2016, consisting of about 7000 undergraduate and 1500 graduate students. The university offered 38 graduate programs, including five professional doctoral degrees and one Ph.D. degree (Table 1). Programs were offered in a variety of modality including on-campus, online, and in mixed mode. In fall 2016, 63% of the graduate students were part-time (37% full-time), 29% were enrolled in classes that were completely online, 63% were female (37% male), and 80% paid in-state tuition (20% out-of-state). The largest programs of enrollment were physical therapy, nursing, business administration, and computer science. The university received an average of 2600 applications per year for admission into one of the graduate programs, accepted an average of 1400 (54% acceptance rate), and enrolled an average of 750 new students each year (54% yield).

Although the university has being having consistently strong and growing graduate enrollment since early 2000s, the enrollment growth leveled off and declined somewhat in the last two years. This was primarily due to a decline in the number of international applicants and recent news regarding a crisis in the institution's city that garnered international attention. Enrolment had dipped just above 8000 students in 2016 from a high of over 8500 in 2014. Further, the university was facing shrinking resources and stronger competition by other higher

Table 1. Programs and majors.

Program/Major Combination	Major Code(s)	College/School	Count	Percent
Accounting (MSA)	ACTG	Management	103	2.2%
Anesthesia	ANE	Health Professions and Studies	143	3.1%
Applied Communication (MA)	ACOM	Arts and Sciences	45	1.0%
Arts Administration (MA)	ARTA	Rackham	41	0.9%
Biology (MS)	BIO	Arts and Sciences	48	1.0%
Business Administration (MBA)	BUS	Management	448	9.7%
Computer Science & Info. Systems (MS)	CAIS	Arts and Sciences	1195	25.9%
Early Childhood Education (MA)	ECHD	Education and Human Services	106	2.3%
Education (Ed. D. & Ed. S.)	EDU	Education and Human Services	224	4.9%
Education (MA)	EDU	Education and Human Services	251	5.4%
English (MA)	ENGL	Arts and Sciences	56	1.2%
Health Education (MS)	HED	Health Professions and Studies	37	0.8%
Liberal Studies (MA)	LBS	Rackham	48	1.0%
Mathematics (MA)	MTH	Arts and Sciences	38	0.8%
Non-Degree	0000	Arts and Sciences	202	4.4%
Nursing (DNP, entry-level)	NUR	Health Professions and Studies	556	12.1%
Physical Therapy (entry-level DPT)	PTP	Health Professions and Studies	367	8.0%
Physical Therapy (transitional DPT)	PTPP	Health Professions and Studies	167	3.6%
Public Administration (MPA)	PUB	Rackham	269	5.8%
Public Health (MPH)	PHS	Health Professions and Studies	198	4.3%
Social Sciences (MA)	SOSC	Arts and Sciences	73	1.6%
		TOTAL	**4615**	**100.0%**

education institutions, including a prevalence of online modality. The university had established a strategic enrollment plan with an explicit goal of increasing graduate enrollment, requiring improvement in the conversion of admitted students to matriculation. There was an urgent need to improve the graduate application "yield" using more intentional and data-driven recruitment strategies and practice.

Dataset

The dataset consisted of 4615 de-identified application records of 3877 unique individuals, submitted over the period spring term 2014 through and including winter term 2017. Table 2 presents the list of variables that were originally developed and were used to collect the data. Values of some of the variables such as distance to the university and the number of days-to-admit-decision were calculated from the respective fields in the original dataset. We also obtained students' scholarship, grant, fellowship, loans, and expected family contribution (EFC) data from a separate information system and merged them with the application record dataset.

Table 2. List of variables.

Variable	Symbol	Description	Statistics
Enter Year	YEAR	Year for which applied	2014 = 163; 2015 = 1612; 2016 = 1504; 2017 = 1336
Enter Term	TERM	Term for which applied	fall = 2833; winter = 1281; spring = 339; summer = 162
Aid Year Code	AIDCODE	Financial aid year code	2014 = 163; 2015 = 1612; 2016 = 1504; 2017 = 1336
Application Number	APPLNO	The application number	1 = 2606; 2 = 968; 3 = 444; 4 or more = 597
Application Date	APPDATE	Date of first application	N/A
Admit Date	ADMITDATE	Date of admission	N/A
Days to Admit	DAYSTOADM	Calculated from the application date to admission decision	Min = 0; Max = 1185 ; Mean = 70.2; St. Dev. = 79.3
Admit Code	ADMTCODE	Type of admission	Conditional = 1673; Probationary = 305; Readmit = 17; Standard = 2620
Student Type	STUTYP	They of students	Continuing (C) = 24; Guest (G) = 35; New (N) = 4144; Readmit (R) = 212; Non-candidate (S) = 200
Primary Program	PRIPGM	Applicant's primary graduate program	See Table 1
Primary Major	PRIMAJOR	Applicant's primary major of interest	See Table 1
Primary Concentration	PRICONC	Applicant's primary concentration of interest	N/A
Primary College	PRICOL	Primary college code	N/A
Residency Code	RESDCODE	Applicant's residency status	Resident = 2502; Non-resident = 2113
State	STATECODE	Applicant's mailing address state	MI = 2712; other U.S. states = 378; International = 1525
Zip Code	ZIPCODE	Applicant's mailing address zip code	N/A
County	COUNTY	Applicant's mailing address county	Genesee = 878; other = 2212; International = 1525
Nation	NATION	Applicant's mailing address country	N/A
Citizenship	CITIZENSHIP	Applicant's citizenship	U.S. = 2843; other countries = 1772
Gender	GENDER	Applicant's gender	Female = 2524; Male = 2061; None = 30
Ethnicity	ETHNICITY	Applicant's ethnicity	Am. Indian = 22; Asian = 157; Black = 407; Hispanic = 86; Non-res. = 1525; White = 2191 ; other = 227
International	INTCODE	Y for international applicant, N for domestic	Y = 1525; N = 3090
GPA	GPA	Applicant's grade point average	Min = 1.60; Max = 4.00; Mean = 3.46; St. Dev. = 0.40
Deposit	DEPOSIT	Yes for applicant's with deposits	Yes = 319; No = 4296
Age	AGE	Age of applicant	Min = 19; Max = 72; Mean = 31.4; St. Dev. = 9.7
Distance	DISTANCE	Distance of residence to the university	Min = 0; Max = 4450; Mean = 127.8; St. Dev. = 340.1
Previous Degree	PREVDEGREE	Level of applicant's previous degree earned	N/A
Education Level	EDULEVEL	Applicant's highest education level	Associate = 32; Bachelors = 2933; Masters = 742; Post-grad = 57; Doctoral = 130; missing = 721
GRE Score verbal	GREVERB	Applicant's verbal score on GRE	N = 1256; Min. = 130; Max = 169; Mean = 145.4; St. Dev. = 8.3
GRE Score quantitative	GREQUANT	Applicant's quantitative score on GRE	N = 1255; Min. = 133; Max = 168; Mean = 150.23; St. Dev. = 5.9

Continued

GRE Score Writing	GREWRITE	Applicant's writing score on GRE	N = 1094; Min. = 1; Max = 6; Mean = 3.2; St. Dev. = 0.9
GMAT Total Score	GMATTOTAL	Applicant's total score on GMAT	N = 316; Min. = 260; Max = 740; Mean = 520.6; St. Dev. = 77.7
GMAT Verbal	GMATVERB	Applicant's verbal score on GMAT	N = 315; Min. = 11; Max = 45; Mean = 27.9; St. Dev. = 7.6
GMAT Quantitative	GMATQUANT	Applicant's quantitative score on GMAT	N = 315; Min. = 8; Max = 51; Mean = 33.8; St. Dev. = 8.0
GMAT Writing	GMATWRITE	Applicant's writing score on GMAT	N = 288; Min. = 1; Max = 6; Mean = 4.6; St. Dev. = 0.9
IELTS Total	IELTS	Applicant's total score on IELTS	N = 866; Min. = 5; Max = 9; Mean = 6.1; St. Dev. = 0.6
TOEFL Overall	TOEFL	Applicant's total score on TOEFL	N = 229; Min. = 45; Max = 112; Mean = 87.2; St. Dev. = 12.1
Year Fin. Aid offer	YRFAOFFER	Total financial aid offer for the aid year	N = 352; Min. $200; Max. = $30,771; Mean = $2608.10; St. Dev. = $3749.37
Term Fin. Aid offer	TRMFAOFFER	Total financial aid offer for the first term	N = 230; Min. $120; Max. = $12,478; Mean = $890.30; St. Dev. = $1487.10
Year Loan offer	YRLOANOFF	Loan amount offer for the aid year	N = 1538; Min. $0; Max. = $41,262; Mean = $18386.33; St. Dev. = $6069.87
Term Loan offer	TRMLOANOFF	Loan amount offer for the aid year	N = 1445; Min. $177; Max. = $20,631; Mean = $10060.95; St. Dev. = $2632.16
EFC	EFC	Expected Family Contribution	N = 1180; Min = $1; Max = $162,441; Mean = $12820.15; St. Dev. = $13896.66
Registered Applied Term	REGINTERM	Registered for the term of admission	No = 2031; Yes = 2584

A preliminary review of the dataset revealed that: 1) not all of the applicants had received an offer (or been awarded) scholarship, grant, and/or fellowship; 2) the vast majority of applicants had received an offer of just one of these types of financial aids; and 3) for the purpose of our study, the effects of scholarship, fellowship and grants were considered to be the same. Hence, we added the amounts of the three types of financial aids and used the total, referred to it as "Financial Aid" instead. For this variable, we recorded the values for the amounts offered for the entire year as well as the first term of enrollment.

4. Methodology

We used Classification and Regression Trees (CART), also known as Decision Tree analysis, to develop the predictive model. CART is an iterative form of data analysis, designed to predict the class of an object based on the values of a set of predictor variables [11]. In each iteration the method chooses a predictor based upon a *tree growing* criterion and divides the remaining objects into two or more groups (tree nodes), each having a different value (or set of values) for that predictor. The process continues until a *stopping criterion* is reached. Some stopping criteria include: predictors are exhausted; the remaining predictors do not have statistically significant predictive powers; maximum number of tree nodes is reached; and/or maximum tree level is reached.

The goal of our model was to determine the predictors that significantly influenced the probability that an applicant would accept the offer of admission

and that he/she would register in the term for which he/she had been admitted. Although some researchers [7] [12] [13] [14] [15] have used logistic regression to develop predictive models for forecasting enrollment, logistic regression is not appropriate for predicting graduate enrollment because not all applicants are subject to the same admission standards [4]. Further, decision trees offer several distinct advantages over logistic regression for our type of modeling. Logistic regression is based upon a priori model, assuming a linear relationship between predictors and the dependent variable, whereas decision tree analysis essentially partitions the dataset into sub-spaces without a prior assumption. Decision tree analysis is not sensitive to a possible non-linear relationship between a predictor and the dependent variable. Decision trees are also more appropriate for models that include several categorical variable with relatively large number of categories. Lastly, interpretation of the results of decision trees is easier and the results are more intuitive.

Our study consisted of three parts. We first used CART to develop a predictive model for all of the applicants, using split sampling for validation. We then used Bayesian Network (BN) to reaffirm the outcome of the CART analysis and compare its predictive power with that of the Decision Tree model. The third part consisted of using the superior technique to develop predictive models for selected sample of the academic majors. This part of the study illustrated nuances that exist when trying to recruit students interested in specific majors.

We used the SPSS Version 22 Decision Tree procedure. The *Chi-square automatic interaction detection (CHAID)* option was chosen with the following parameters: maximum tree depth = 15; minimum cases in parent node = 40; and minimum cases in child node = 10. The CHAID option was selected because it allowed for multi-level node splitting rather than just binary splitting [16]. The values of parameters were selected based upon our extensive experience with decision tree analysis and to prevent the model from "overfitting" the data [17]. The dependent variable was registered in Term. For validation, we used split sampling with 80% for the training sample and 20% for the test sample. Not all of the graduate programs under consideration required GMAT or GRE exams. Variables representing GMAT and GRE exams were excluded from the first part of our study with the goal of improving the overall yield of graduate applications. Preliminary analysis of the data revealed that State Code had the highest discriminating power and was selected as the first predictor for splitting. However, for this predictor there were too many split levels with rather small sub-populations. We therefore recoded State Code into three categories: Michigan plus its adjacent states of Ohio, Indiana, and Illinois; Other U.S. states; and International.

Figure 1 represents the top three levels of the decision tree associated with the training sample. The test sample tree was identical to the one for training sample but with different counts. The training sample consisted of 3711 records with 56.3% Registered in Term and 43.7% not registered. The test sample had 904 records with 54.5% registered in term and 45.5% not registered. The resulting

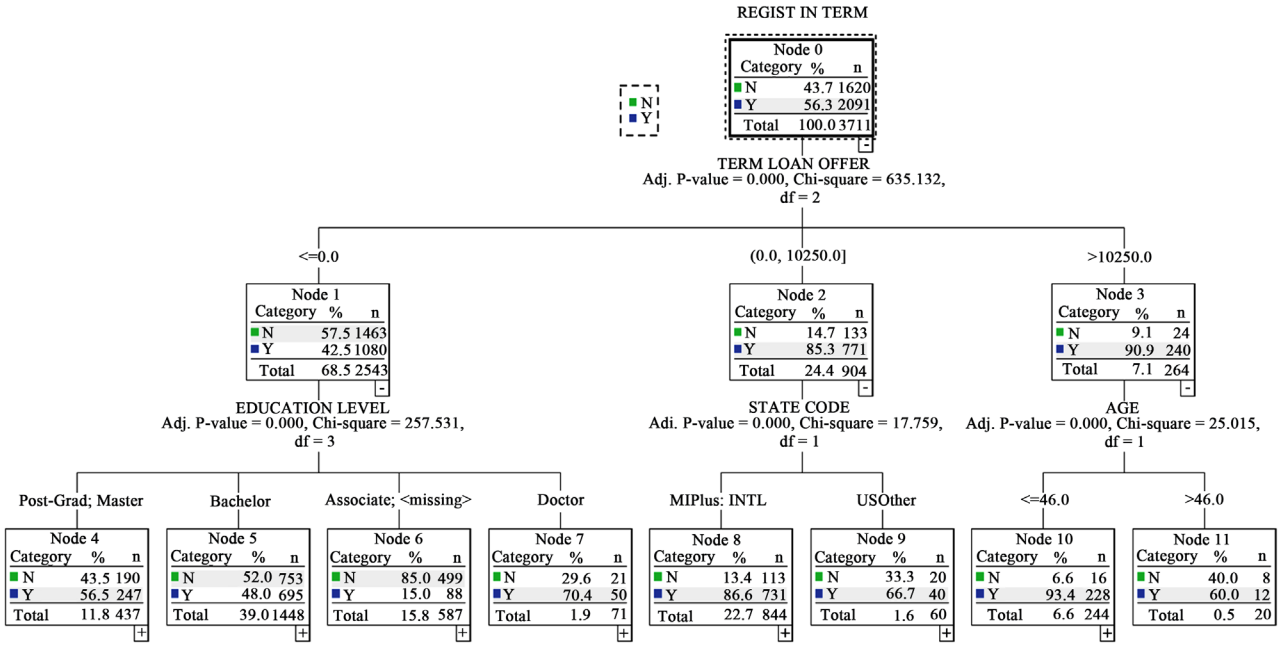

Figure 1. Decision tree for all applications, top three levels.

training decision tree had 125 nodes, including 68 terminal nodes and 11 levels. The overall correct classification (prediction power) of the training sample was 79.0% and that of the test sample was 74.3%. The resulting risk estimates were 21.0% and 25.7% for the training sample and test sample, respectively. The relative closeness of the prediction powers of the training sample and test sample signified a rather robust mode.

The most discriminating predictor having the first level split was the Term Loan Offer with three split levels: less than or equal to zero (Node 1); Greater than zero but less than or equal to $10,250 (Node 2); and greater than $10,250 (Node 3). The split with the highest percentage of Registered in Term was the applicants with Term Loan Offer greater than $10,250 (Node 3) with 90.9% registered and the split with lowest figure was applicants with zero Term Loan Offer at 42.5% registered. The significantly higher percentage of registered applicants with positive Term Loan Offers in Nodes 2 and 3 can be used by the admission office to develop strategies for enhancing the chances that an applicant would accept an offer of admission and would actually enroll. The question then becomes how much term loan or other type of financial aid should be offered to increase the registered in term by a given percentage. The answer to this question can be determined by running a crosstab consisting of registered in term versus term loan offer amount for the domestic students.

Additional enrollment strategies can be developed by examining lower levels of the decision tree. For instance, the sub-tree below Node 1, applicants with zero Term Loan Offer was formed by splitting Education Level (Figure 2). The split levels consisted of Post-Grad and Master (Node 4), Bachelor (Node 5), Associate and missing (Node 6), and Doctor (Node 7). The highest percentage of

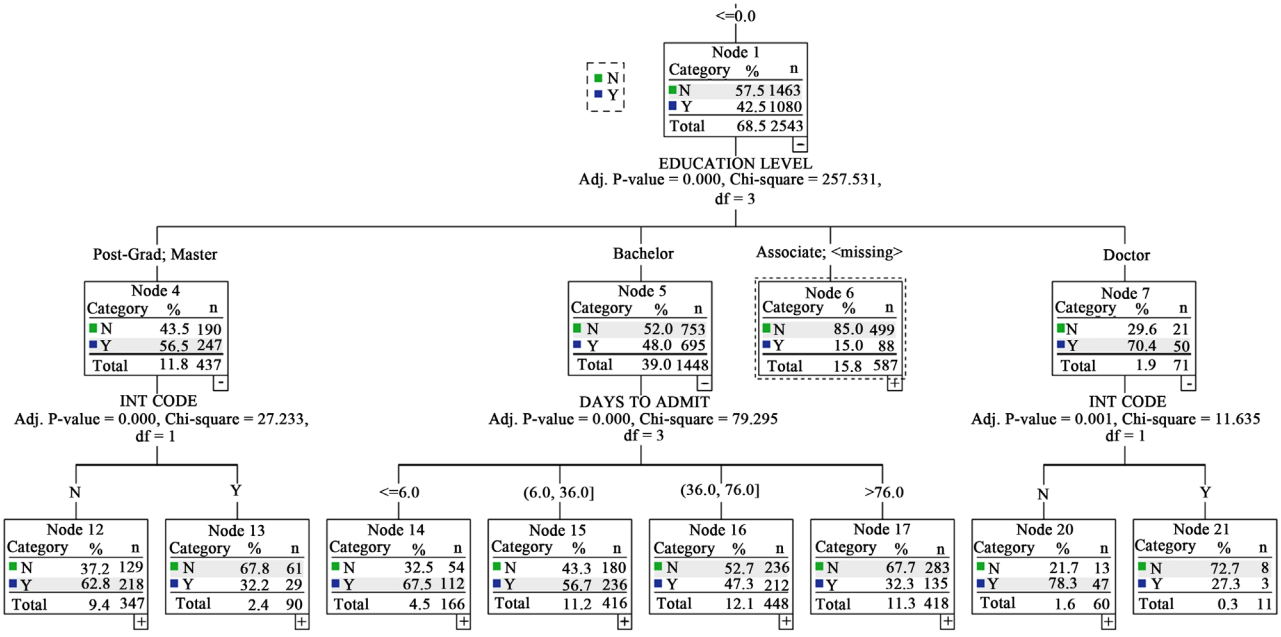

Figure 2. Sub-Tree for domestic students with zero term loan offer, top three levels.

registered in term was associated with applicants with doctoral degree (Node 7) at 70.4% and the applicants with lowest registered in term were those with an associate degree or missing value at 15.0%.

Another interesting discovery is the sub-tree below the applicants with Bachelor degrees formed by splitting Days To Admit at levels: six days or less; more than six days but less than or equal to 36 days; more than 36 days but less than or equal to 76 days; and greater than 76 days. The percent registered in term varies from 32.3% for those with admission decision (Days To Admit) taking longer than 76 days and those with admission decision made in six days or less at 67.5%, more than double the former rate. This is another example of potential use of decision tree methodology in enhancing recruitment strategies. The admission office can encourage the graduate program faculty and administrators to decrease the time it takes to make admission decisions.

Comparison with Bayesian Network Model

We compared the results of the above decision tree analysis with those of a Bayesian Network (BN) model to reaffirm the predictive power of the decision tree technique. A BN is a directed acyclic graph with nodes representing the variables (both dependent and independent) and the edges representing possible dependencies between the end nodes of each edge [18]. For a given BN, we can compute the conditional probability of one node, given the observed values of the other nodes. A BN therefore can be used as a predictive model where interest lies in determining the conditional (posterior) probabilities of the values of the dependent variable (called the class node) for a given set of values of the independent variables.

For this part of the analysis we used *Knostanz Information Miner* (*Knime*)

Analytic Platform Version 3.3, a comprehensive open solution data analytic package developed in Zurich, Switzerland [19]. To make the comparison as similar as possible, we used both the Naïve Bayesian model and the Decision Tree model within *Knime* (see Figure 3). We used the same training and test dataset that we had used in the decision tree analysis for developing and testing the two *Knime* models. The training sample had 3711 records and the test sample had 904 records.

The Bayesian model *learner* maximum number of distinct categories per categorical variable was set at 20. When executed, the *learner* excluded College Code, County Code, Primary Concentration, Primary Major, Primary Program, Previous Degree, and Citizenship because these predictors had too many categories. The model also removed Deposit because of too many missing values (most programs of study at the study institution do not require an enrollment deposit). Registered in Term was used as the output variable with 2091 "Y" count and 1620 "N" count. Table 3 presents the list of variables that were included in the Bayesian network model. The model overall prediction rate for the test sample was 70.7% sensitivity of 78.7% and specificity of 61.1%.

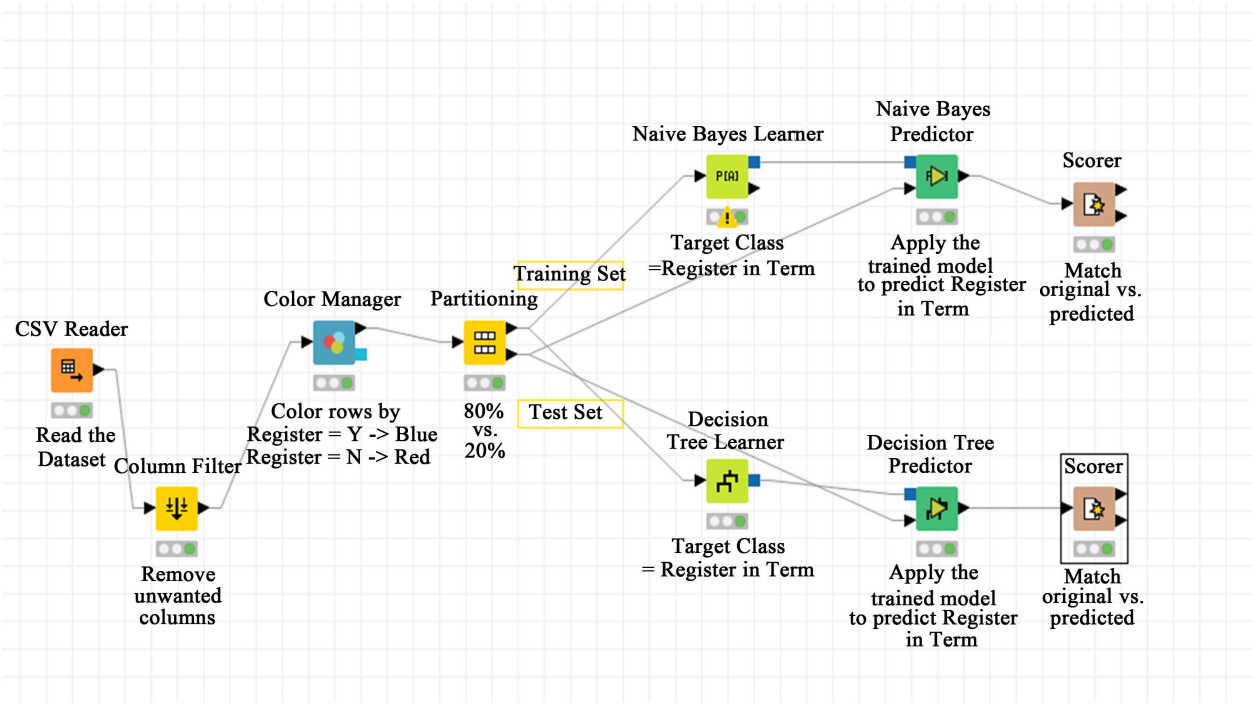

Figure 3. The Knime model.

Table 3. Bayesian model variables.

Enter Year	Application Code	Residency Code	GPA	TOEFL	Term Loan Offer
Enter Term	Days to Admit	Gender	Age	Year FA Offer	EFC
Aid Year Code	Admit Code	Ethnicity	Distance to University	Term FA Offer	Register in Term
Application Number	Student Type	International Code	Education Level	Year Loan Offer	State Code

The Decision Tree model resulted in a tree with 67 nodes, including 30 terminal nodes, and 11 levels. The first level predictor was Term Loan Offer with split levels less than or equal to $93 with 42.5% registered in term and greater than $93 with 86.6% registered in term. The model overall prediction rate for the test sample was 74.4% with sensitivity of 78.3% and specificity of 69.8%. Although Knime uses a different node splitting algorithm from that of SPSS, there were similarities in the structure of the two trees. For instance, in the Knime tree predictors such as Days to Admit, Citizenship and Primary Major were among the top level predictors. Comparing the overall prediction rates of the Bayesian network analysis and the Decision Tree methodology indicates that for our dataset, the Decision Tree approach resulted in higher prediction power with significantly better specificity.

Examining Specific Majors

The above analyses focused on developing predictive models for the overall applicants to the graduate programs. In this section, we examined the predictors that might significantly impact enrollment in specific programs. It is conceivable that characteristics of applicants interested in different academic disciplines might vary significantly. For instance, an applicant interested in a health professions program could be very different than a person interested in a business program or humanities program. Further, admission requirements for different programs differ significantly. For instance, the business program required the GMAT score whereas the computer science program required GRE score and the nursing programs required different levels of education (associate's, bachelor's, or master's, depending on the specific program). Our goal here was to establish a framework whereby a graduate admission office would become sensitive to possible nuances that might exist for applicants to different disciplines rather than trying to create a predictive mode for each and every academic program. We focused on the four most populous programs in our dataset to illustrate the concept. They included Computer Science & Information Systems, Nursing, Business Administration (including the certificate and accounting programs), and Physical Therapy programs.

We used the decision tree analysis for the programs as well. Because the datasets for the majors were smaller than that for all of the programs, we set the minimum number of cases per parent node at 20, minimum number of cases per child node at five, and the maximum tree depth at 10. The validation method was split sampling with an 80/20 split for the training set and test set, respectively.

There were a total of 1195 applicants for the computer science programs with 31.5% registered in term and 68.5% not registered. The relatively high percentage of not registered was due to a high number of international applicants (approximately 90.1%), with only 28.4% registered. The resulting decision tree had 68 nodes, including 37 terminal nodes and seven levels. The overall correct classification for the training set was 81.0% and that of the test set was 72.5%. **Figure 4** represents the top three levels of decision tree for the computer science majors.

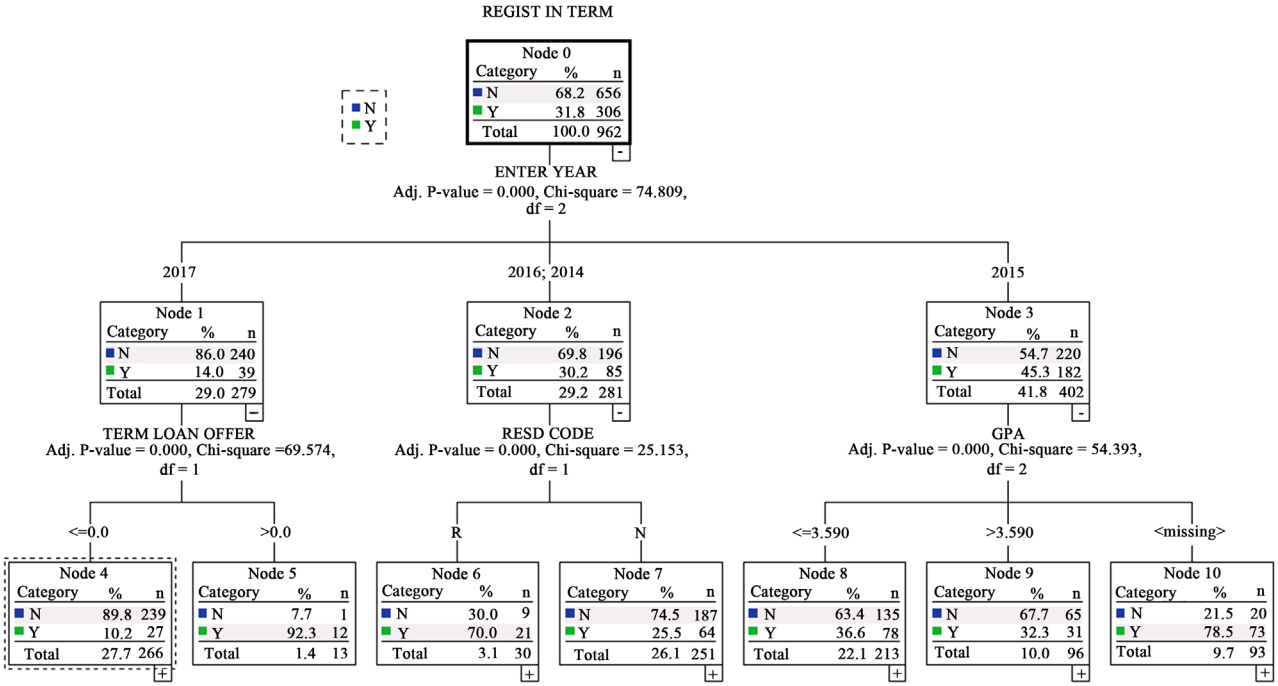

Figure 4. Decision tree for computer science programs, top three levels.

The first level predictor was Enter Year, followed by Term Loan Offer, Residency Code, and GPA. Examining the 2017 applicants, there is a significant difference in Registered in Term between those with no loan offer at only 10.2% and those with a positive loan offer at 92.3%. For the 2014 and 2016 applicants, there is somewhat similar distinction between resident applicants at 70.0% Registered in Term versus non-residents at 25.5%.

There were 556 applicants for the nursing programs, with 73.2% registered in term and 26.8% not registered. The nursing majors' decision tree had 27 nodes, including 14 terminal nodes and five levels. Figure 5 represents the top three levels of the decision tree. The overall correct classification for the training set was 84.1% and that of the test set was 72.4. The top level predictor was Term Loan Offer with two split levels: less than or equal to zero (Node 1) with 55.2% registered in term; and greater than zero (Node 2) with 89.1% registered in term. The significant difference in percentage of registered in term for these two groups of applicants clearly delineates the effect of financial aid in accepting the offer of admission and registering in the program.

The second level predictor below Node 1 was Enrollment Deposit. Those with no deposit registered in term at 48.2% and applicants with deposit registered at 78.6%. This information is extremely useful for highly selective academic programs with limited capacity. Such programs often over-admit students to ensure that they would fill their target cohort to capacity. The knowledge of the percentage of applicants who have made their enrollment deposit can be used to develop a more reliable enrollment forecast. The predictor below Node 2 was Days to admit with split levels less than or equal to 140 days at 92.3% registered in

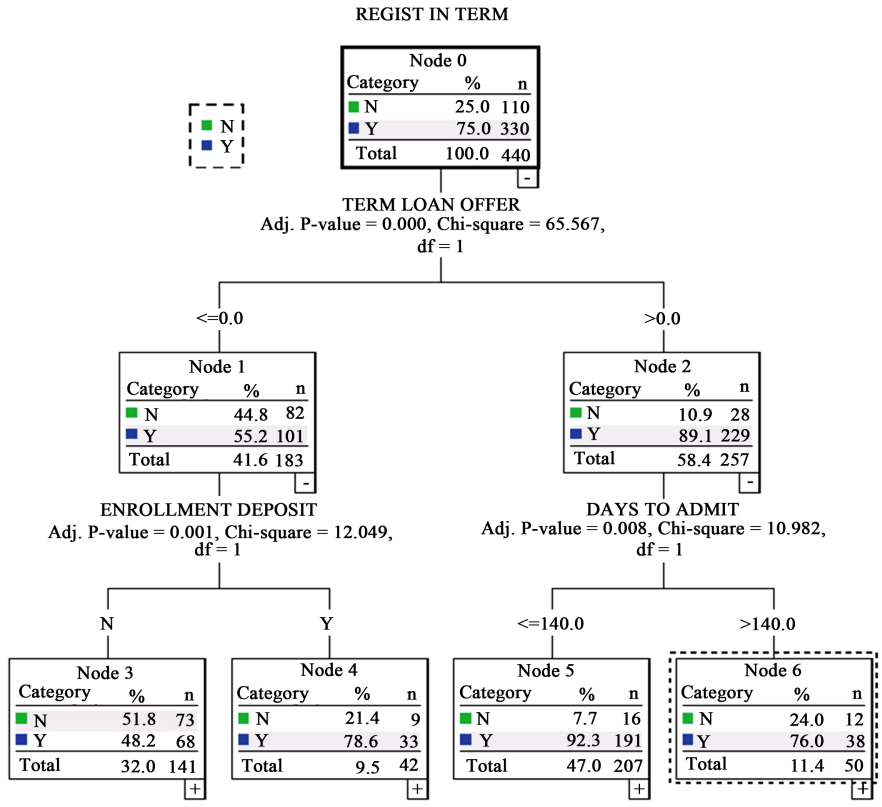

Figure 5. Decision tree for the nursing majors, top three levels.

term and more than 140 days at 76.0% registered in term. This information can be shared with admission officers to show the potential impact of taking too long to make admission decisions.

There were a total of 549 applicants for the Master of Business Administration, Master of Science in Accounting, and the business certificate program with 65.4% registered in term and 34.6% not registered. The resulting decision tree (**Figure 6**) had 32 nodes, including 17 terminal nodes, and seven levels. The overall prediction rate for the training set was 79.9% and that of the test set was 64.9%. The first level predictor was State Code with left node (Node 1) consisting of International and other U.S. states applicants with 42.1% registered in term and right node (Node 2) representing Michigan Plus applicants with 76.4% registered.

The above information can be used to develop a more precise enrollment forecast. That is, when using a yield rate of admitted applicants in computing the enrollment forecast, rather than using a fixed yield rate, one could weigh the number of applicants from Michigan Plus states higher than applicants from international and other states. The second level split below Node 1 was formed by splitting Enrollment Deposit. Applicants with no deposit had 38.5% registered in term and those with deposit had 88.9% registered in term. The second level split below Node 2 was formed by splitting Term Loan Offer. Applicants with less than or equal to zero loan had 68.7% registered in term and those with greater

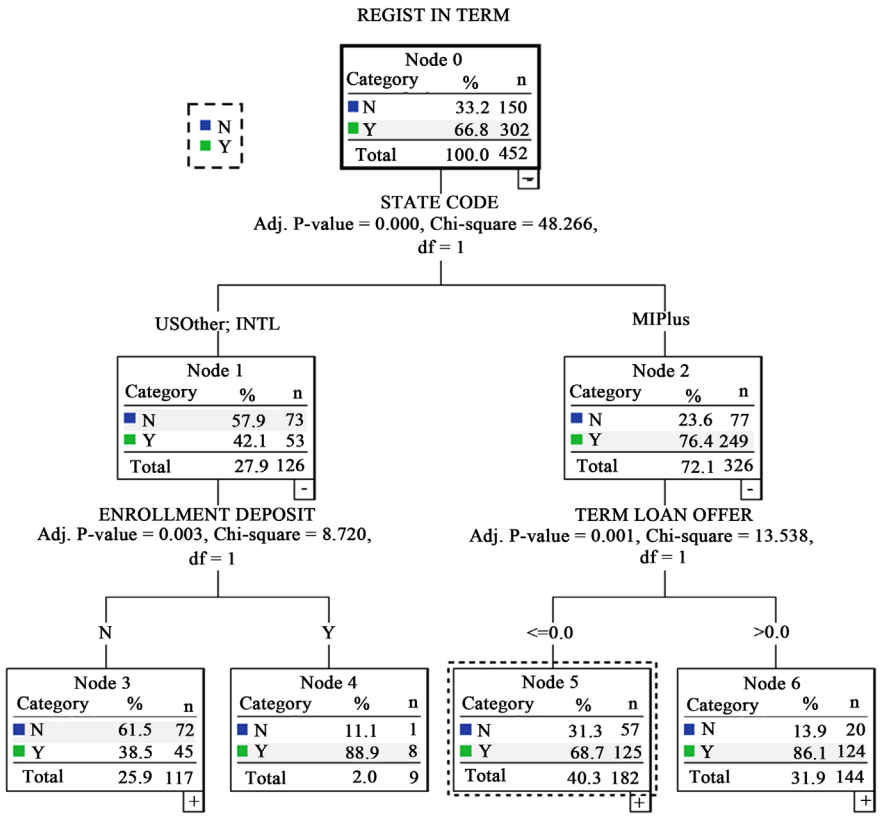

Figure 6. Decision tree for business majors–top three levels.

than zero had 86.1% registered in term. As with the State Code, the Enrollment Deposit and Term Loan Offer can be used to arrive at more precise enrollment forecasts for the business majors.

There were 534 applicants for the physical therapy programs, with 51.9% registered in term and 48.1% not registered. The resulting decision tree had 27 nodes, including 15 terminal nodes and five levels (Figure 7). The overall correct classification for the training set was 84.6% and that of the test set was 87.6%. The top level predictor was Year Loan Offer with two splits: less than or equal to zero with 32.7% registered in term (Node 1), and greater than zero with 90.9% registered in term (Node 2). Interestingly, the second level predictor for both nodes was Primary Major with splits of PTPP and PTP but with different percent registered in term. For instance, the PTP (entry-level program) majors with a positive Year Loan Offer register at significantly higher rate (93.2%) compared to PTPP (post-professional program) majors (42.9%) with positive Year Loan Offer.

5. Discussion

This study involved developing predictive models for assessing the likelihood that a graduate applicant would enroll in a program of study during the semester following admission decision. The models were based upon actual application information of over 4600 graduate applicants at a mid-sized public university

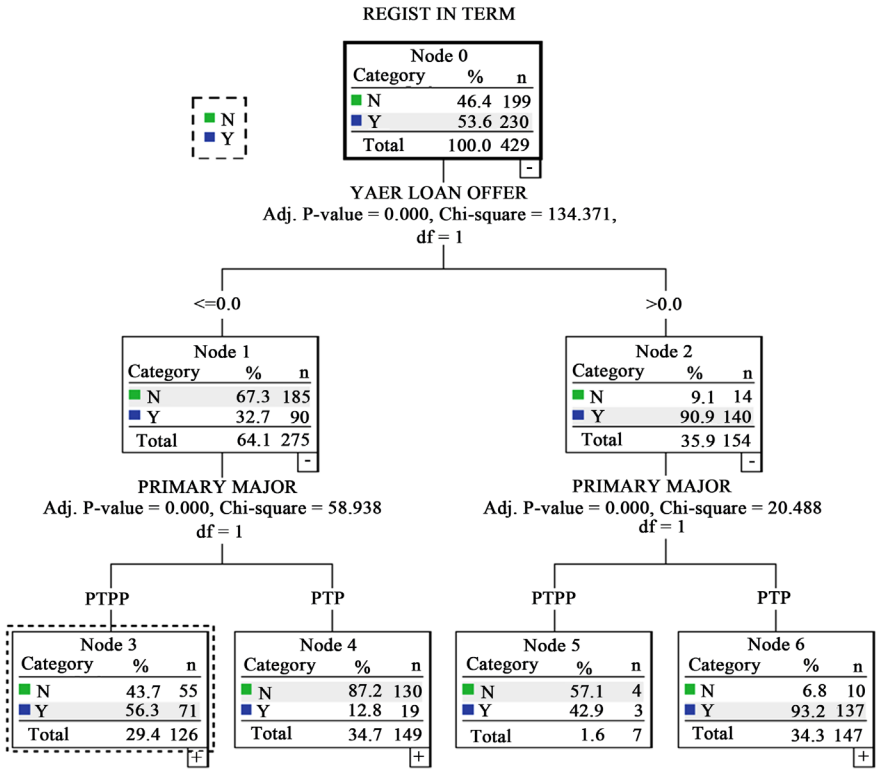

Figure 7. Decision tree for physical therapy majors–top three levels.

over a three-year period. The applicants' dataset included application information such as demographic characteristics, test scores, financial aid information, and other pertinent data. The first part of the study consisted of developing a predictive model using Decision Tree analysis for all applicants, irrespective of their academic major of interest. We then compared the Decision Tree model's performance with that of a Bayesian Network model to reaffirm its validity and predictive power. The Decision Tree-based model out-performed the Bayesian model for our dataset. The third part of the study involved using Decision Tree methodology to develop predictive models for a sample of four popular academic majors. The trees were used to illustrate more precise enrollment forecasting and recruitment strategies for overall recruiting efforts as well as possible strategies for the sample majors.

A major contribution of this study to the strategic enrollment management literature pertains to the development of predictive modeling for graduate applicants. Graduate students can be an essential and even a critical component of a university strategic enrollment plan for institutions that offer graduate education. Accordingly, it is vital that such a plan utilizes data-driven and more advanced modeling techniques in forecasting graduate enrollment. Unlike undergraduate applicants who face almost the same admission standards for a given university, graduate applicants must satisfy institutional requirements such as minimum grade point average (GPA) and English language proficiency as well as programmatic requirements such as aptitude tests or professional license.

Another important contribution of this study is the establishment that factors which influence an applicant to enroll in a graduate program of study might vary by academic discipline. Hence, recruitment efforts, targeting potential graduate student populations should incorporate elements designed to appeal to the overall population of students as well as components designed to target specific majors.

The study is limited since our predictive model did not include qualitative and subjective factors such as reputation of the university or program rankings. This limitation can be addressed by surveying the applicants before or after they enroll and then try to incorporate their responses into the predictive models. However, such an approach could be susceptible to possible flaws in applicants' recollection if done after enrollment and potential to influence their opinion if done before the admission decision is made. Another limitation of the study is with respect to its population of applicants associated from a mid-sized public institution. Applicants at much larger universities with numerous academic disciplines might exhibit different dynamics with respect to factors that influence their decision to accept an offer of admission and enroll. Also, applicants at private universities could behave differently than those of public institutions. Nonetheless, we have presented a framework for developing predictive models that can be implemented at other types of institutions using their own historical data.

References

[1] Bransberger, P. and Michelau, D.K. (2016) Knocking at the College Door. Western Interstate Commission for Higher Education (WICHE), Boulder, CO. http://knocking.wiche.edu/

[2] Selingo, J. (2013) Colleges Struggling to Stay Afloat. New York Times. http://www.nytimes.com/2013/04/14/education/edlife/many-colleges-and-universities-face-financial-problems.html

[3] Mitchel, M., Leachman, M. and Masterson, K. (2016) Funding down, Tuition up, State Cuts to Higher Education Threaten Quality and Affordability. Center on Budget and Policy Priorities, Washington, DC. https://www.cbpp.org/sites/default/files/atoms/files/5-19-16sfp.pdf

[4] Langston, R., Wyant, R. and Scheid, J. (2016) Strategic Enrollment Management for Chief Enrollment Officers: Practical Use of Statistical and Mathematical Data in Forecasting First Year and Transfer College Enrollment. *Strategic Enrollment Management Quarterly*, **4**, 74-89. https://doi.org/10.1002/sem3.20085

[5] Hossler, D. and Bontrager, B. (2014) Handbook of Strategic Enrollment Management. Jossy-Bass, San Francisco, California.

[6] Thomas, E., Dawes, W. and Reznik, G. (2001) Using Predictive Modeling to Target Student Recruitment: Theory and Practice. AIR Professional File, 78. Association for Institutional Research, Tallahassee, FL.

[7] DesJardins, S.L. (2002) An Analytic Strategy to Assist Institutional Recruitment and Marketing Efforts. *Research in Higher Education*, **43**, 531-553. https://doi.org/10.1023/A:1020162014548

[8] Goenner, C.F. and Pauls, K. (2006) A Predictive Model of Inquiry to Enrollment. *Research in Higher education*, **47**, 935-956.

https://doi.org/10.1007/s11162-006-9021-8

[9] Ledesma, R. (2009) Predictive Modeling of Enrollment Yield for a Small Private College. *Atlantic Economic Journal*, **37**, 323-324. https://doi.org/10.1007/s11293-009-9177-7

[10] Shrestha, R.M., Orgun, M.A. and Busch, P. (2016) Offer Acceptance Prediction of Academic Placement. *Neural Computing and Applications*, **27**, 2351-2368. https://doi.org/10.1007/s00521-015-2085-7

[11] Faraway, J. (2006) Extending the Linear Models with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models. Chapman and Hall/CRC Texts in Statistical Science, Boca Raton, Florida; London, UK; New York, NY.

[12] Paulsen, M.B. (1990) College Choice: Understanding Student Enrollment Behavior. ASHE-ERIC Higher Education Reports, Report No. 6, George Washington University, School of Education and Human Development, Washington DC.

[13] Berge, D.A. and Hendel, D.D. (2003) Using Logistic Regression to Guide Enrollment Management at a Public Regional University. AIR Professional File, 86. Association for Institutional Research, Tallahassee, FL.

[14] Ahluwalia, P.M.S. (2006) Enrollment Prediction Using Bayesian Multiple Logistic Regression (Order No. 1435459). ProQuest Dissertations & Theses Global, 304918817.

[15] Metcalfe, Y.L. (2012) A Logistic Regression and Discriminant Function Analysis of Enrollment Characteristics of Student Veterans with and without Disabilities (Order No. 3523362). Military Database, ProQuest Dissertations & Theses Global, Technology Collection, 1038155485.

[16] Shmueli, G. (2007) Classification Trees: CART vs. CHAID. Business Analytics, Statistics, Teaching. http://www.bzst.com/2006/10/classification-trees-cart-vs-chaid.html

[17] Bramer, M. (2013) Avoiding Overfitting of Decision Trees. In: Principles of Data Mining. Undergraduate Topics in Computer Science. Springer, London, United Kingdom. https://doi.org/10.1007/978-1-4471-4884-5_9

[18] Pearl, J. (1988)Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufman Publishers, Inc., San Francisco, California.

[19] Berthold, M.R., Cebron, R., Dill, F., Gabriel, T.R., *et al.* (2009) KNIME: The Konstanz Information Miner: Version 2.0 and beyond. SIGKDD Explore Newsletter.