A MODEL TO PREDICT MATRICULATION OF CONCORDIA COLLEGE APPLICANTS

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Kaylin Marie Pavlik

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Program:
Applied Statistics

May 2017

Fargo, North Dakota

North Dakota State University
Graduate School

Title

A MODEL TO PREDICT MATRICULATION OF CONCORDIA COLLEGE APPLICANTS

By

Kaylin Marie Pavlik

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Rhonda Magel
Chair

Dr. Seung Won Hyun

Dr. Simone Ludwig

Approved:

May 1, 2017
Date

Dr. Rhonda Magel
Department Chair

ABSTRACT

Colleges and universities are under mounting pressure to meet enrollment goals in the face of declining college attendance. Insight into student-level probability of enrollment, as well as the identification of features relevant in student enrollment decisions, would assist in the allocation of marketing and recruitment resources and the development of future yield programs.

A logistic regression model was fit to predict which applicants will ultimately matriculate (enroll) at Concordia College. Demographic, geodemographic and behavioral features were used to build a logistic regression model to assign probability of enrollment to each applicant. Behaviors indicating interest (campus visits, submitting a deposit) and residing in a zip code with high alumni density were found to be strong predictors of matriculation. The model was fit to minimize false negative rate, which was limited to 18.1 percent, compared to 50-60 percent reported by comparable studies. Overall, the model was 80.13 percent accurate.

ACKNOWLEDGEMENTS

Thank you to my advisor Dr. Rhonda Magel, who helped me visualize and set the foundation for this study before I even enrolled in the statistics program. Without her guidance and expertise, this thesis would not have been possible.

I would also like to thank the individuals who formed the committee to review this thesis, Dr. Seung Won Hyun and Dr. Simone Ludwig.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION AND LITERATURE REVIEW

College admission is becoming more and more competitive; not just for the students, but for the institutions themselves. According to a survey conducted by the *Chronicle of Higher Education*, 40 percent of small private colleges missed their enrollment and tuition revenue goals in 2016 (Hoover and Lipka, 2016).

Soaring student debt levels and the pressure to validate the worth of a college degree have contributed to the differentiation of higher education beyond traditional colleges and universities. Traditional offerings are becoming less relevant as two-year, certificate, and for-profit institutions address the growing need for "middle skill," often technical, jobs: positions that require more than a high school diploma but not a complete bachelor's degree. Demand for these type of workers is expected to increase as baby boomers retire (Holzer and Lerman, 2008).

A decline in student enrollment has direct consequences for institutional operations through the resulting reduction in tuition revenue. Enrollment managers are under more and more pressure to keep or grow enrollment in order to maintain tuition revenue. As institutions tighten their budgets and pressure to hit enrollment goals intensifies, strategic enrollment management is more important than ever.

Some colleges endeavor to increase declining enrollment by marketing to larger populations of potential students to increase applicants. Other strategies aim to strengthen the percentage of admitted students who enroll (the yield rate) or enhance student retention rates.

Many colleges are now exploring predictive modeling to assist in any or all of these strategies. There is little academic research into this topic (Thomas et. al., 2001), particularly regarding private liberal arts schools. Numerous consulting firms practice predictive enrollment modelling, though most keep their practices proprietary, leaving much to be learned.

This study aims to leverage enrollment data to build a model to predict which admitted students will enroll (matriculate) at Concordia College in Moorhead, Minnesota. The objective of the model is to (1) generate student-level probability of enrollment to assist in allocation of financial aid and recruitment funds and (2) describe relevant influential factors in student enrollment decisions to aid in future yield-focused policies and programs.

1.1. Student college choice

Many researchers have studied and developed models of student college choice. Jackson (1982) combined sociological and economic models to outline three stages: preference, exclusion and evaluation. In the first phase, students develop a preference or attitude toward college education influenced by their family, peers, personal goals and academic achievement. In the second phase, students exclude certain options to develop their choice set of potential institutions. In the final phase, students evaluate and make a choice.

Hossler and Gallagher (1987) outlined a similar three-phase model divided into predisposition, search and choice stages. In their model, search is a longer phase encompassing both the exclusion and the evaluation phases of Jackson's (1982) model. It includes the search activities students use to find colleges as well as the search activities colleges use to find students.

In the preference or predisposition phase, students must decide to pursue higher education instead of an alternative career path. Background characteristics like socioeconomic status have been shown to impact college enrollment, with students of higher socioeconomic status four times more likely to attend college as those of lower status (Peters, 1977). The same study found quality of high school curriculum and attending a high-status high school are positively related to college enrollment (Peters, 1977). Parental and peer influence has also been

shown to affect college attitudes as well as specific college choice (Conklin, 1981; Tillery, 1973).

The development of a choice set of potential institutions is likewise impacted by demographic and relational characteristics. High-ability students tend to conduct more sophisticated searches earlier on (Litten, 1983) and are more likely to expand their geographic search radius farther than lower-ability students (Zemsky, 1983).

In the final decision, the perception of quality is an important factor in the inclusion of an institution as a first or second choice (Jackson and Chapman, 1984). Financial aid positively influences student choice toward an institution, except for high-income, no-need students (Freeman, 1984).

1.2. Institution-specific studies

Several studies have been conducted that predict the application or enrollment behavior of students at particular institutions. Application behavior occurs during the search phase of college choice, while enrollment behavior occurs during the choice phase. These studies were able to demonstrate that assigning enrollment probabilities to each student can be used to effectively segment students to strategically target marketing and recruitment resources.

DesJardins (1999) studied ACT survey data to describe the application behavior of students to a large research university. Their linear regression model found academic features (test scores, high school rank), proximity to the institution, and family income to be important predictors of application behavior.

Goenner and Pauls (2006) fit a regression model to predict which inquiries (students who have indicated interest but have not yet submitted an application) would ultimately enroll at the University of North Dakota. The model's overall accuracy was 89 percent. While its ability to

correctly predict those who would not enroll was 97 percent, its ability to correctly predict matriculants was only 36 percent.

Roth (2007) used academic and demographic predictors to build several linear regression models to predict enrollment of applicants to the Ohio State University. The study yielded an overall accuracy of 60 percent, likely due to the limitation imposed on the study in the exclusion of behavior and geographic variables.

Thomas (2001) built a predictive model to determine which admitted students would enroll at the State University of New York in Stony Brook. Demographic, academic, geographic and behavioral variables were used to fit a regression model that reached overall accuracy of 70 percent. The model was better at predicting which students would not enroll (82 percent accurate) rather than students who would enroll (46 percent).

DesJardins (2002) aimed to predict which admitted students at a large university would enroll using application and ACT survey data to fit a logistic regression model. The model achieved an overall accuracy of 65.7 percent and found geographic features, high school size and historical yield, and application date to be significant predictors of enrollment.

2. METHODOLOGY

Concordia College is a four-year private liberal arts college of the Evangelical Lutheran Church in America offering more than 50 majors, including 15 honors majors and 12 preprofessional programs. Central selling points for Concordia are small, engaged classes and integrative learning; global curriculum and programs; and student involvement, particularly in music and athletics.

The aim of this study is to provide student-level probability of enrollment for first year domestic students admitted to Concordia. Institutional goals to increase enrollment yield support the decision to focus on the enrollment decision of students already admitted to the college, rather than on the application behavior of prospective students. Furthermore, much more information is available about admitted students, leading to better prediction and improved understanding of the underlying factors that influence enrollment.

2.1. The recruiting cycle

It is helpful to describe a full recruiting cycle in order to give context to the model. See Figure 1 for a high-level overview of the recruiting timeline.

Communication with prospective students may begin as early as their sophomore year in high school. Many institutions engage with vendors to conduct search campaigns to gauge and identify interested students at this point. Communication with younger students focuses on visiting campus and introduction to institutional features, such as program offerings, student life, outcomes and core values.

High school seniors may apply at any time from July 1 through the first day of class due to Concordia's rolling admission approach. Recruitment efforts focus on application generation throughout the fall. During this time, admission representatives present at college fairs and visit

high schools; the college hosts multiple large group visit events; and marketing communications focus on visiting campus and submitting an application.

During the winter, focus shifts to converting admitted students to enrolled students, often through conversations about scholarships and aid. Academic (merit) and performance (music, theatre, visual arts, and speech) scholarship competitions are held on campus November through February. Aid packages are typically awarded starting in early February. Aid packages cannot be awarded until the student's FAFSA (Free Application for Federal Student Aid) has been submitted to the college. In the years of this study, FAFSA submission opened January 1.

Admitted student events are held on campus in March and April. During the same time period, admission reps contact each admitted student in their territory to discuss their options involving enrollment decisions, program choice and financial aid.

Admitted students can "save their spot" in the incoming class by placing an enrollment deposit of $300. As mandated by the "Statement of Principles and Good Practice" written by the National Association for College Admission Counseling (NACAC), these deposits are fully refundable until May 1, a date also widely known as National Decision Day. Typically, 70 to 80 percent of the final freshman class submitted their enrollment deposit prior to May 1.

The full recruitment cycle for a cohort ends on the tenth day of classes, at which time a final enrollment number is recorded.

Application generation						Admit conversion (yield)								
July	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**	**Jan.**	**Feb.**	**March**	**April**	**May**	**June**	**July**	**Aug.**	**Sept.**
Application opens						FAFSA opens		Model run		Decision day				Classes begin

Figure 1. Recruiting cycle timeline.

In order to be useful in practice, student-level probabilities of enrollment would have the most utility if delivered by March 1 of the recruiting year. March 1 was selected to maximize the data available about student behaviors (e.g., FAFSA submission or scholarship competition participation) while still allowing for adequate reaction time to allocate resources and make decisions regarding final recruiting methods.

2.2. The statistical model

When the outcome variable is binary (enrolled/did not enroll), logistic regression is the most appropriate model (Abraham and Ledolter, 2006). The model is specified as

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

where P_i is the probability that student *i* will enroll at Concordia; X_i is a vector of demographic characteristics, Y_i is a vector of geodemographic characteristics, and Z_i is a vector of behavioral intent characteristics; α, β_i, δ_i, γ_i are estimated coefficients; and ε_i is a normally distributed random error term. The predicted variable is the log of the odds that student *i* will enroll at Concordia. The model is estimated using the iteratively reweighted least squares (IWLS) method using R, open-source statistical software, which is the default of the function glm using the binomial family (Abraham and Ledolter, 2006).

2.3. The data

Three cohorts will be used in this analysis. The first two, students admitted for enrollment to the fall 2014 and fall 2015 terms, have served as the model development set (N=3539) for the testing on the third cohort (N=2341), students admitted for enrollment for the fall 2016 term.

The data is restricted to domestic first year admitted students (that is, international students, transfer students and students denied admission are not included). The primary source

for the data is the Office of Admission, though the Advancement Office provided alumni location information and geodemographic information was pulled from the US Census (2014).

2.4. Predictive variables

The variables included in the model can be broken into three categories: demographic, geodemographic, and behavioral. The variables have been chosen to reflect factors of interest due to anecdotal support and specific institutional qualities presumed to be points of difference in student college choice.

Because the model is designed to predict which admits will ultimately enroll as of March 1, all data has been limited to information in hand as of that date. If a student completed a campus visit on March 15 or submitted their FAFSA on April 1, that information would not be included in this model as it would not have been available at the time of prediction.

Demographic characteristics include gender, ethnicity, religious preference, legacy relationships, ACT score, grade point average (GPA), and music or athletic interest. One third of students participate in a music ensemble and 25 percent of the student body participate in one of the college's 22 NCAA Division III athletic teams. The ability to participate in college-level athletics or music performance at a degree that doesn't eclipse academics is often used as a selling point in conversations with prospective students, and thus the reason why they are included for consideration in this model.

Geodemographic characteristics includes information about the student's zip code of residence: distance from Moorhead, population, percent of population that is white, median income, alumni density and peer admitted student density. While residents across a certain zip code are surely not homogeneous, a "neighborhood effect" has been shown to affect educational choices (Leventhal and Brooks-Gunn, 2002). As students search for communities in which they

feel comfortable and can picture themselves in, the percentage of non-white students or students with more family financial aid need may be a draw or a detractor. Alumni and peer admitted student density are included to incorporate the influence of peers and community members on college choice (Tillery, 1973).

Behavioral characteristics include number of campus visits, FAFSA submission, invitation and acceptance into the honors program, participation in a performance scholarship competition, and participation in an academic scholarship competition by March 1. These actions are assumed to indicate continued student interest and therefore should predict intent to enroll.

The response variable is a discrete variable representing whether or not the student ultimately matriculated (enrolled in fall classes) at Concordia. The variables are described in Table 1, and the descriptive statistics for the training set are given in Table 2.

Table 1. Definitions of the variables.

Response variable	
Enrolled	1 for enrolled in fall classes, 0 otherwise
Demographic characteristics	
Gender	1 for male, 0 for female
Ethnicity	1 for non-white, 0 otherwise
Religious indicator	1 for any religion listed, 0 otherwise
Lutheran indicator	1 for Lutheran, 0 otherwise
Legacy relationship	1 for any legacy relationship, 0 otherwise
ACT	Standardized test score
GPA	Grade point average
Music interest	1 for any music activity reported as co-curricular or academic interest, 0 otherwise
Athletic interest	1 for any athletic activity reported as co-curricular activity or interest, 0 otherwise
Geodemographic characteristics	
Distance	Miles from Moorhead
Zip code population	2014 US Census, in thousands
Zip code percent white	2014 US Census
Zip code median income	2014 US Census, in thousands
Zip code alumni density	Ratio of alumni to population in a given zip code
Zip code peer admitted student density	Ratio of students admitted to same term to population in a given zip code
Behavioral (actions completed prior to March 1)	
Campus visits	Number of regular, non-event campus visits
FAFSA	1 for Free Application for Federal Student Aid (FAFSA) submitted, 0 otherwise
Honors program invitation	1 for invited to honors program, 0 otherwise
Honors program accepted	1 for accepted/enrolled in honors program, 0 otherwise
Performance scholarship competitor	1 for participated in on-campus competition for performance scholarship, 0 otherwise
Academic scholarship competitor	1 for participated in on-campus competition for academic scholarship, 0 otherwise
Deposited	1 for enrollment deposit paid, 0 otherwise

Table 2. Descriptive statistics of the training set.

Variable	N	%	Mean	S.E.	Min.	Max.
Enrolled	961	26.64				
Demographic characteristics						
Gender	1221	37.66				
Ethnicity	284	8.76				
Religious indicator	1084	33.44				
Lutheran indicator	547	16.87				
Legacy relationship	365	11.25				
ACT	3242	-	25.45	0.07	13	36
GPA	3242	-	3.61	0.01	1.63	4.0
Music interest	764	23.57				
Athletic interest	756	23.32				
Geodemographic characteristics						
Distance	3242	-	257.33	7.50	0	6,797
Zip code population	3242	-	20.23	0.27	0.08	113.45
Zip code percent white	3242	-	89.65	0.21	1.76	100.00
Zip code median income	3242	-	61.51	0.31	21.92	146.70
Zip code alumni density	3242	-	0.47	0.01	0	2.55
Zip code peer admitted student density	3242	-	0.06	0.00	0	1.23
Behavioral (actions completed prior to March 1)						
Campus visits	3242	-	0.74	0.01	0	7
FAFSA	1126	34.73				
Honors program invitation	173	5.33				
Honors program accepted	64	1.97				
Performance scholarship competitor	343	10.58				
Academic scholarship competitor	578	17.83				
Deposited	731	22.54				

The model will return a numeric prediction for each student ranging from 0 to 1. A cutoff point will need to be chosen in order to assign each student a prediction of 1 ("will enroll"), or 0 ("will not enroll). In previous studies, cutoff points of 0.13 (Goenner, 2005), 0.30 (Thomas et. al, 2001), and 0.36 (Roth, 2008) have been chosen to reflect that the binary outcome is not represented equally in the dataset. Because "enrolled" students account for only 26.64 percent of the training set, a cutoff point near 0.26 may be supported.

The primary goal will be to minimize the false negative rate while maximizing the true positive rate. In the case of student enrollment, while correctly identifying which students will enroll is clearly a top priority, incorrectly classifying a student as "will not enroll" could have costly effects.

Previous studies have found it much easier to accurately predict students who will not enroll (true negatives) at the expense of the accuracy of predicting those who will (true positives). Goenner (2005) was able to reach a 97 percent true negative rate (specificity) but only 36 percent true positive rate (sensitivity); Thomas et. al (2001) hit 82 percent specificity but only 42 percent sensitivity. Particular emphasis to balance specificity with sensitivity will be given in the cutoff point decision in this model.

To determine the true positive and negative rates, the model will be run on the test set and a confusion matrix created to compare model predictions with the actual fall 2016 results.

3. THE RESULTS

3.1. Model assumptions

Logistic regression does not have the same restrictive assumptions of linear regression. While linear regression requires the variables to be linearly related and multivariate normal, and the residuals need to be normally distributed and exhibit homoscedasticity, logistic regression carries none of those restrictions.

Logistic regression does assume that the variables are not collinear and that the sample size contain at least 10-30 cases per parameter estimated.

No variables exhibit correlation greater than 0.75. The most strongly correlated independent variables are the religious indicator and whether the student is religiously affiliated with the Lutheran faith at 0.64. See Figure 2 for a correlation plot of all values.

The sample size assumption is met at N=3242 in the model development set. The development set is large enough to allocate 100 observations per parameter (22 independent variables plus an intercept term).

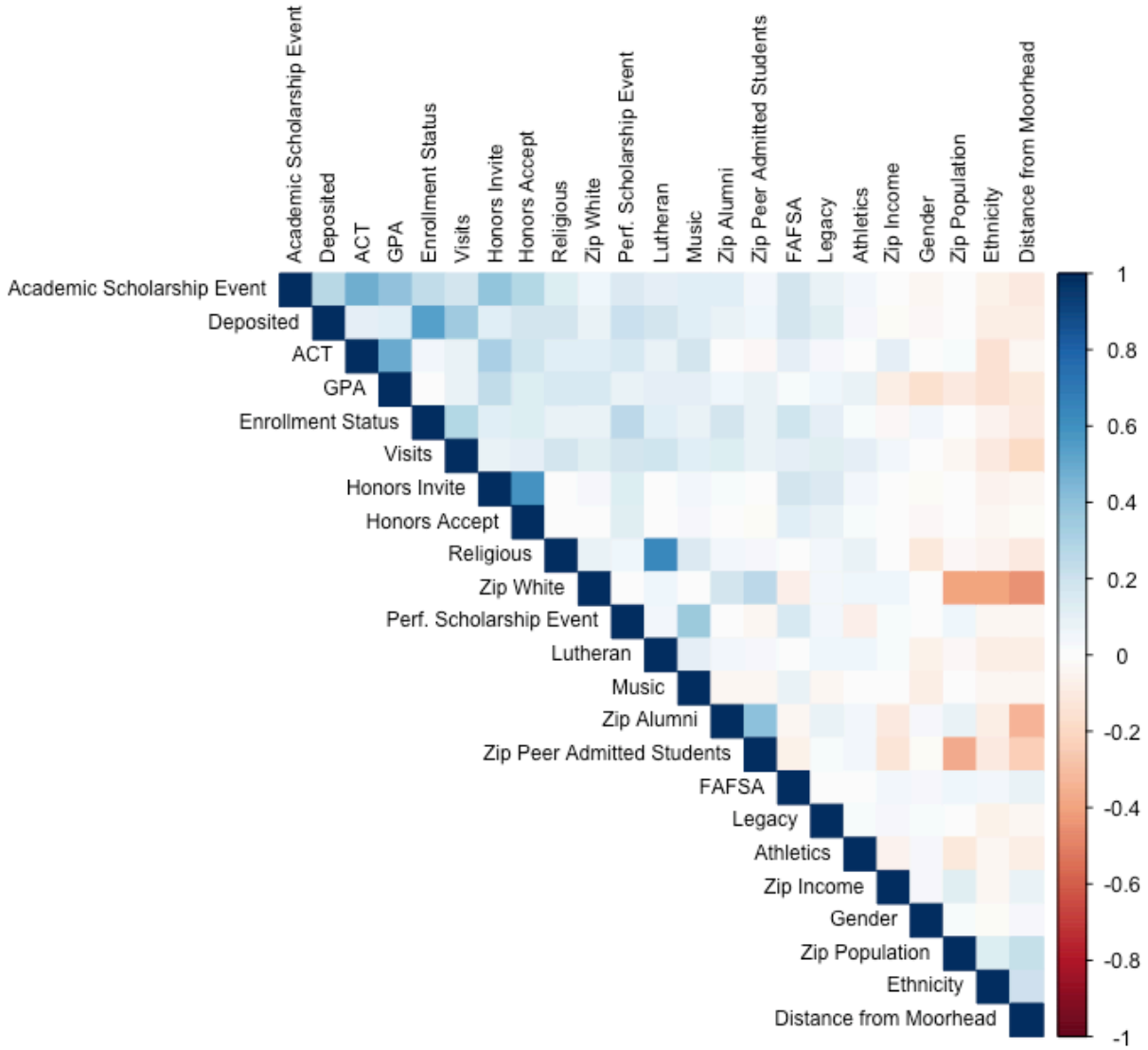

Figure 2. Correlation matrix plot.

3.2. Fitted model

Table 3 contains the logistic regression results produced using the training set. The table contains the full model results and the final restricted model. The final model was chosen using stepwise regression to minimize model AIC (Abraham and Ledolter, 2006). Smaller AIC (Akaike's information criterion) is preferred.

Table 3. Training set results.

	Final model			Full model		
Variable	**Odds ratio**	**Estimate**	**P value**	**Odds ratio**	**Estimate**	**P value**
Intercept		0.8109	0.2633	1.8439	0.6119	0.4573
Demographic characteristics						
Gender	1.2869	0.2523	0.0128	1.2658	0.2357	0.0214
Ethnicity				1.1600	0.1484	0.4429
Religious indicator				0.8578	-0.1533	0.2659
Lutheran indicator				1.2850	0.2508	0.1313
Legacy relationship				1.1593	0.1478	0.3333
ACT	0.9358	-0.0664	0.0000	0.9395	-0.0624	0.0001
GPA	0.4284	-0.8478	0.0000	0.4272	-0.8504	0.0000
Music interest				0.8563	-0.1551	0.2260
Athletic interest				1.0135	0.0134	0.9085
Geodemographic characteristics						
Distance				1.0001	0.0001	0.4589
Zip code population				0.9992	-0.0008	0.8425
Zip code percent white	1.0162	0.0161	0.0020	1.0173	0.0171	0.0048
Zip code median income	0.9951	-0.0049	0.0951	0.9950	-0.0051	0.0920
Zip code alumni density	1.7806	0.5769	0.0000	1.7663	0.5690	0.0000
Zip code peer admitted student density				1.5582	0.4425	0.5245
Behavioral (actions completed prior to March 1)						
Campus visits	1.3594	0.3071	0.0000	1.3571	0.3053	0.0000
FAFSA	1.6328	0.4903	0.0000	1.6279	0.4873	0.0000
Honors program invitation				0.9494	-0.0519	0.8529
Honors program accepted				1.2206	0.1993	0.6307
Performance scholarship competitor	3.9749	1.3800	0.0000	4.3648	1.4738	0.0000
Academic scholarship competitor	3.1976	1.1624	0.0000	3.1718	1.1543	0.0000

Deposited	11.1539	2.4118	0.0000	11.1176	2.4085	0.0000
Model AIC			2740.5			2755.6
Null Deviance			3941.0			3941.9
Residual Deviance			2716.5			2709.6

Demographic characteristics found to significantly impact the odds of enrollment include ACT score and GPA, both of which represent lesser odds as scores increase. For each point increase in ACT score, the odds of enrollment decrease by 6.5 percent. For each decimal increase in GPA (e.g. from 2.5 to 3.5), the odds of enrollment decrease by 57 percent.

Median income, percentage of inhabitants that are white, and alumni density by zip code were geodemographic characteristics found to be significant. For each $1,000 increase in median income, the odds of enrollment decrease by 0.5 percent, though for each percent increase in alumni density and white inhabitants, odds of enrollment were 1.01 and 1.78 times higher.

Unsurprisingly, the most influential behavioral variable was whether or not the student made an enrollment deposit. While many students place deposits at multiple schools and request refunds after making a final decision, placing a deposit before March 1 increases the odds of enrollment more than 11 times.

Participation in an on-campus scholarship competition increased a student's odds of enrollment 3.97 times for performance scholarships (art, theatre, music, speech) and 3.20 times for academic scholarships.

Also predictive of enrollment were the submission of a FAFSA, which indicates a student's intent to request financial aid from Concordia, and the number of non-special-event campus visits attended. Submitting a FAFSA increased the odds of enrollment 1.63 times, and each additional visit to campus increased the odds of enrollment 1.36 times.

3.3. Model significance

To assess the model, the deviance of the null and final model were computed. The deviance assesses the fit of a given model versus the fit of the fully saturated model (where each data point has its own parameter, resulting in *n* parameters; Abraham and Ledolter, 2006).

The null deviance, or deviance with only the intercept, is 3940.99; the residual deviance, or deviance including the parameters, is 2716.55. The difference, 1224.44, is the test statistic, which follows a chi-square distribution with degrees of freedom equal to the number of variables, 11 (p-value = 0.0000). The model is significant.

3.4. Test set prediction

In order to test the accuracy of the model, we held out applicants for the fall 2016 term. The cutoff point 0.20 was chosen in order to bound the false negative rate at 18.1 percent. The model had an overall accuracy of 80.13 percent. It was able to accurately predict 81.6 percent of students who would ultimately enroll at Concordia, while also accurately predicting 79.6 percent of students who would not ultimately enroll.

Table 4. Confusion matrix.

	Actual not enrolled	Actual enrolled	Total
Predict not enrolled	1383 *(79.6%)*	90 *(18.1%)*	1473
Predict enrolled	354 *(20.4%)*	407 *(81.9%)*	761
Total	1737 *(100%)*	497 *(100%)*	

The receiver operating characteristic (ROC) curve is a visual expression of the trade-offs between true positive rate (the y-axis) and the false positive rate (the x-axis). The area under the

curve (AUC) represents the strength of the model. The closer the curve comes to the dashed 45-degree line, the less accurate the test. The area under the curve for this model is 0.8076.

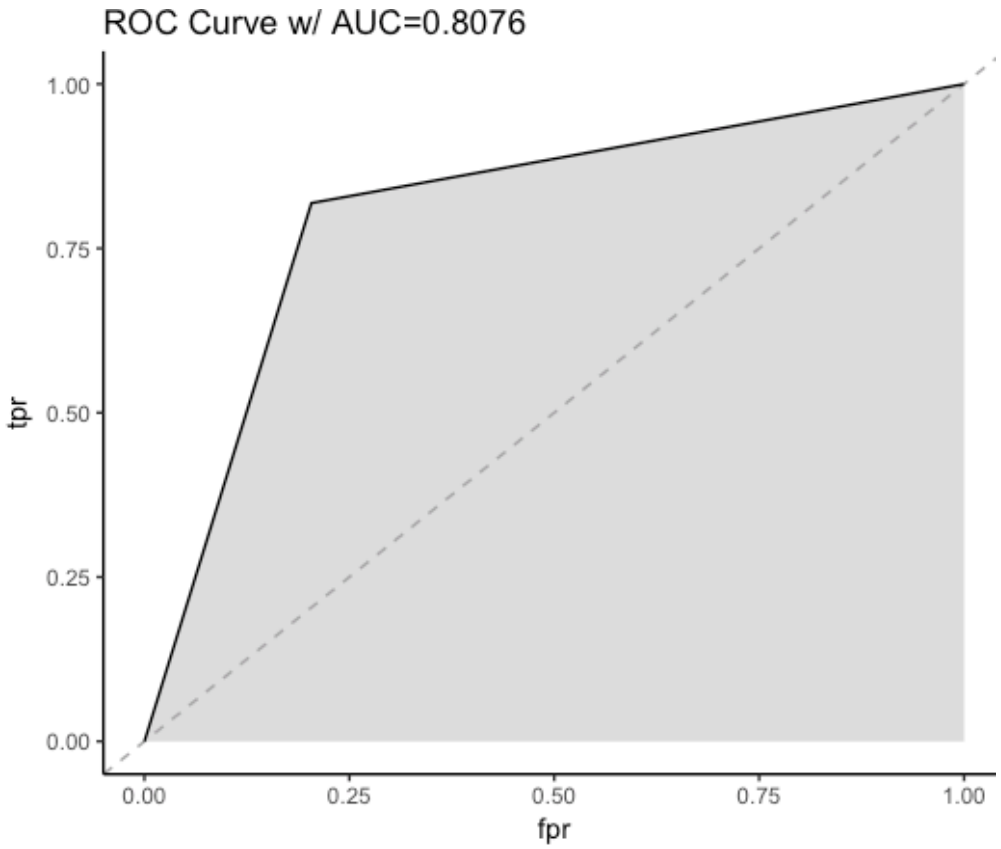

Figure 3. Receiver operating characteristic curve plot.

3.5. Limitations

Financial aid package amounts and family need figures are noticeably omitted from this study. Prior to 2017, the FAFSA could not be submitted until the student's parents submitted their taxes for the prior year. Because of this, many aid packages had not yet been awarded by March 1. In 2017, the U.S. Department of Education changed the regulations to permit submission using taxes from two years back ("prior prior taxes") and allowed submission as early as October 1. The effects of this change are not yet known. Future models may be able to incorporate financial aid package information to better predict student enrollment status.

Additionally, no two enrollment years are the same, which may mandate the redevelopment of this model in as little as one to two years. Enrollment marketing methods (such as name-buying, new geographic market development, changing vendor agreements, and staff turnover) lead to inconsistencies between years. Moreover, each graduating class of high school seniors is unique and poses fresh challenges and opportunities.

4. IMPLICATIONS AND DISCUSSION

The fitted model has several implications for future recruiting and marketing efforts. It confirmed many assumptions already held about religion, diversity, the role of alumni, campus visits and scholarships, while challenging ideas about gender and high-ability students.

Concordia is a college of the Evangelical Lutheran Church in America and faith and tradition play a large role in academic affairs and student life. As such, enrollment managers generally expect students of Lutheran faith to be more likely to identify with Concordia. The model assigned an odds ratio of 1.28 to the feature; that Lutheran students would be 1.28 times more likely to enroll is consistent with expectations, though the feature was not found to be significant ($p=0.1313$). Future models may consider including this feature since it is near the cutoff.

Campus diversity is an institutional priority for Concordia as well as many private colleges both regionally and nationally. According to U.S. Department of Education's College Scorecard (http://collegescorecard.ed.gov), four-year private colleges in Minnesota range from 55 percent (Augsburg College) to 85 percent (Bethel University) white, with Concordia at the top end with 84 percent white students. A working group, council, Office of Diversity and new positions have been created to support institutional goals of increasing diversity and enhancing inclusion on campus. The finding that students from zip codes with higher percentages of white inhabitants are more likely to enroll, and conversely that students from more diverse zip codes are less likely, supports the need for better campus resources and support for diverse students.

Institutions often cite an alumni network of professionals and successful graduates as a benefit of their degrees, crafting a fictional future scenario where shared alma mater with the interviewer means the difference between an offer and being passed over for a job. The visible

presence of these alumni in a student's hometown may make this scenario seem all the more likely. The density of alumni in a student's zip code (the ratio of alumni to total population) has a positive effect on enrollment, with each one percent increase of alumni increasing the odds of enrollment by 78 percent. Recruitment efforts could benefit from geographic targeting and market development in alumni-dense areas.

Visits to campus play a large role in the recruitment process. Visits are highly customizable: baseline visits include a campus tour, dining hall meal and a presentation; once tailored to the student's interests, they may include a band practice, economics class, meeting with a professor or athletic training. It has been shown that students who visit apply and enroll at higher rates than students who don't visit. Moreover, post-visit surveys show that students' opinion of Concordia and reported likelihood to apply increase dramatically after a few hours on campus, so it does not come as a surprise that with each campus visit a given student is 1.35 times more likely to enroll. Getting students to campus will remain a priority in the recruiting cycle.

Scholarship competitions are two-day events on campus that include a banquet, information session and an interview and/or audition. The extended campus visit combined with talent recognition and a high likelihood of additional aid makes these competitions potent affinity-building events. Additionally, students who apply for a competitive scholarship are certainly already more invested in Concordia than those who do not, so it is no surprise that they are 3.97 and 3.19 times (performance and academic scholarship competitors, respectively) more likely to enroll. Future recruitment efforts may focus on expanding these programs to other domains.

Sticker price is a common hurdle for private colleges to overcome with prospective students. For the 2016-2017 academic year, comprehensive fees (tuition, room, and board) for Concordia were $44,764. On average, students received more than $25,000 in aid, bringing the net cost down to less than $20,000. The relative benefits of discounting the full rate or lowering tuition overall have been subjects of debate in recent years. The strategy involving discounting through awarding merit scholarships builds affinity by making the student feel valued and sought after. Students who submitted their FAFSA to Concordia were found to be 1.63 times more likely to enroll than those who didn't. It follows that students who don't need financial assistance are less swayed by the discounting strategy, since the model found that students from zip codes with higher median income were less likely to enroll. "Full pay" students, or students who require no aid, free up aid for students who do need it, so recruitment efforts going forward may pursue alternative methods for pursuing students from affluent communities.

Females account for 59 percent of the student body at Concordia and 62 percent of admitted students in the training set, but the model reports that male admitted students are 1.28 times more likely to enroll than females. This is an interesting quirk that may support marketing campaigns dedicated to growing the male student share.

At Concordia, the average ACT score is 25, with 50 percent scoring between 22 and 28 out of the possible 36 points (Fact Book). In previous years, students scoring more than 28 on the ACT have been segmented out of the general population and marketed to differently with invitations to apply early, and later, to enroll in the academic honors program. This approach may not be paying off; the model shows decreased odds of enrollment as ACT scores and GPAs increase. Moreover, the invitation to and even acceptance of the honors program were not found

to be significant predictors of enrollment. Preferential treatment of high-ability students and the use of the honors program as a recruitment tool may warrant reevaluation going forward.

Unsurprisingly, the biggest predictor of enrollment is the enrollment deposit. Students can "melt" and request a refund prior to the May 1 deadline, but fewer than 20 percent do so. Historical marketing and recruitment efforts focus on moving students from admitted to deposited for this reason. The model supports those efforts, indicating than students who have deposited by March 1 are 11.15 times more likely to enroll come September than those who have not yet deposited.

5. CONCLUSION

This study fit a logistic regression model to enrollment data to identify which admitted students would enroll at Concordia College in Moorhead, Minnesota. The model was designed to be run on March 1 to best maximize completeness of enrollment data as well as leave time for subsequent strategy changes.

Admitted student features including demographic, geodemographic and behavioral characteristics were used as independent variables. The recruitment cohorts for the fall 2014 and 2015 terms were used to build the model, which was tested on the fall 2016 term data and found to be 80.13 percent accurate. A cutoff point of 0.20 was chosen to minimize the false negative rate and resulted in the correct identification of 81.6 percent of students who would enroll and 79.6 percent of students who would not enroll. The ability to identify and target high- and moderate-probability students will assist in the allocation of financial aid and recruitment funds.

The features described by the fitted model will also aid in future yield efforts by helping enrollment managers identify effective programs and recruitment strategies. Students who visit campus, participate in scholarship competitions, submit their FAFSA and deposit are more likely to enroll than those who do not. Students who come from zip codes with higher density of alumni and less diversity are also more likely to enroll, while high-ability students and those from more affluent zip codes are less likely to enroll. This information supports campus diversity initiatives and enhancements to the visit and scholarship competition programs. It also suggests targeting alumni-dense geographic areas and reevaluating the use of the honors program as a recruitment tool.

Enhanced understanding of the factors influencing college choice, particularly for students considering Concordia College, is an asset in the ever-intensifying college admission industry.

6. REFERENCES

Abraham, Bovas and Johannes Ledolter. *Introduction to Regression Modeling.* Thomson Brooks/Cole, 2006, Belmont, CA.

Concordia College Office of Institutional Effectiveness. “Fact Book Part 2: Student Enrollment.” *Concordia College*, 17 Oct. 2016. Intranet. Accessed 14 Mar. 2017.

Conklin, Mary E., and Ann Ricks Dailey. "Does consistency of parental educational encouragement matter for secondary school students?" *Sociology of Education*, 1981, pp. 254-262.

Craft, William J. “Whole Self, Whole Life, Whole World: The Plan for Concordia College, 2012-2017, Report to Campus.” *Concordia College,* 5 Oct. 2012, www.concordiacollege.edu/files/resources/strategicplan.pdf. Accessed 14 Mar. 2017.

DesJardins, Stephen L., Halil Dundar, and Darwin D. Hendel. "Modeling the college application decision process in a land-grant university." *Economics of Education Review,* vol. 18, no. 1, 1999, pp. 117-132.

DesJardins, Stephen L. "An analytic strategy to assist institutional recruitment and marketing efforts." *Research in Higher Education*, vol. 43.5, no. 5, 2002, pp. 531-553.

Goenner, C. and K. Pauls. “A Predictive Model of Inquiry to Enrollment.” *Research in Higher Education*, vol. 47, no. 8, 2006, pp. 935-956.

Freeman, H. “The impact of "no-need" scholarships on the matriculation decision of academically talented students.” *Proceedings of the American Association of Higher Education*, 1984, Chicago.

Holzer, Harry and Robert Lerman. “America’s Forgotten Middle-Skill Jobs.” *Urban Institute*, 18 Mar. 2008, www.urban.org/research/publication/americas-forgotten-middle-skill-jobs. Accessed Mar. 7, 2017.

Hoover, Eric and Sara Lipker. “Enrollment Goals Remain Elusive for Small Colleges.” *The Chronicle of Higher Education,* 11 Dec. 2016, www.chronicle.com/article/Enrollment-Goals-Remain/238624. Accessed 6 Mar. 2017.

Hossler, Don, and Karen S. Gallagher. "Studying student college choice: A three-phase model and the implications for policymakers." *College and university*, vol. 62, no. 3, 1987, pp. 207-21.

Jackson, Gregory A. "Public efficiency and private choice in higher education." *Educational evaluation and policy analysis*, vol. 4, no. 2, 1982, pp. 237-247.

Jackson, R., and Chapman, R. "The influence of no-need aid and other factors on college choices of high ability students." *College Board Annual Forum.* 1984.

Leventhal, Tama, and Jeanne Brooks-Gunn. "The neighborhoods they live in: the effects of neighborhood residence on child and adolescent outcomes." *Psychological Bulletin*, vol. 125, no. 2, 2000, pp. 309.

Litten, Larry H. *Applying market research in college admissions*. New York: College Entrance Examination Board, 1983.

National Association for College Admission Counseling. “Statement of Principles and Good Practice.” *NACAC*, 1 Oct. 2016, nacacnet.org. Accessed 13 Mar. 2017.

Peters, W. B. *Fulfillment of short-term educational plans and continuance in education*. Washington D.C.: National Center for Educational Statistics. 1977.

The R Project for Statistical Computing. "R: A language and environment for statistical computing." R Foundation, 2017, www.R-project.org. Accessed 17 Mar. 2017.

Roth, Sadie E. "A model to predict Ohio University student attrition from admissions and involvement data." Dissertation, Ohio University, 2008.

Thomas, E., W. Dawes, and G. Reznik. "Using Predictive Modeling to Target Student Recruitment: Theory and Practice." *AIR Professional File*, vol. 78, no. 11, 2001.

Tillery, Dale. *Distribution and differentiation of youth; a study of transition from school to college*. Ballinger Pub. Co, 1973.

Zemsky, Robert, and Penney Oedel. *The structure of college choice*. College Entrance Examination Board, 1983.

APPENDIX. R CODE

```
library(corrplot)
library(ROCR)
library(ggplot2)

data <- read.csv("enrollment_data.csv", stringsAsFactors=F)

# test and train sets
train1 <- data[data$Year!="2016 Fall" & data$App=="1", -18]
test1 <- data[data$Year=="2016 Fall" & data$App=="1", -18]

# check for collin
co1 <- round(cor(train1[,-c(2,3)]),4)
co1_concern <- co1[abs(co1)>0.75] # just an identity matrix --> good

# make a correlation plot
co1.df <- train1[,-c(2,3)]
colnames(co1.df) <- c("Enrollment Status", "ACT", "GPA", "Distance
from Moorhead", "Zip Income", "Zip Population", "Zip White", "Zip
Alumni", "Music", "Athletics", "Legacy", "Gender", "Ethnicity",
"Lutheran", "Religious", "Deposited", "FAFSA", "Honors Invite",
"Honors Accept", "Visits", "Perf. Scholarship Event", "Academic
Scholarship Event", "Zip Peer Admitted Students")
co1 <- round(cor(co1.df),4)
par(mar=c(rep(5,4)))
corrplot(co1, method="color", type="upper", tl.col="black",
tl.cex=.75,mar=c(0,0,1,0), order="FPC")

# summary stats of training set
summary(train1)
library(psych)
describe(train1[,-c(2,3)], type=2)

# fit model
fit1g <- glm(Final.Status ~ ., train1[,-c(2,3)], family="binomial")
fit1gs <- step(fit1g, direction="both", trace=0) # stepwise selection
coef1s <- exp(cbind(Odds.Ratio = coef(fit1gs), confint(fit1gs))) # get
odds ratios

# make predictions and evaluate cutoff options
test1$Predictions_raw <- unlist(predict(fit1gs, test1,
type="response"))
test <- data.frame()
for( cutoff in c(.05, .075, .1, .15, .2, .25, .3, .35, .4, .45, .5,
.55, .6, .65, .7, .75, .8, .85, .9)) {
    test1$Predictions <- ifelse(test1$Predictions_raw > cutoff, 1, 0)
    results <- table(test1$Predictions, test1$Final.Status)
```

```
    accuracy <- round((results[1,1] + results[2,2])/sum(results[,1],
results[,2]),2)
    fn <- round(results[1,2]/sum(results[,2]),2)
    tn <- round(results[1,1]/sum(results[,1]),2)
    fp <- round(results[2,1]/sum(results[,1]),2)
    tp <- round(results[2,2]/sum(results[,2]),2)
    row <- data.frame(cutoff, accuracy, tp, tn, fp, fn)
    test <- rbind(test, row)
}

# categorize predictions based on cutoff point
test1$Predictions <- ifelse(test1$Predictions_raw>0.20, 1, 0)

# create ROC plot
pred <- prediction(test1$Predictions, test1$Final.Status)
perf <- performance(pred, "tpr", "fpr")
auc <- performance(pred, measure="auc")
auc <- auc@y.values[[1]]
roc.data <- data.frame(fpr=unlist(perf@x.values),
                       tpr=unlist(perf@y.values),
                       model="GLM")
ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +
geom_ribbon(alpha=0.2) + geom_abline(intercept=0, slope=1,
col="darkgrey", lty=2) + geom_line(aes(y=tpr)) + labs(title=paste0("ROC
Curve w/ AUC=", round(auc,4))) + theme_classic()

# test for overall model significance
D <- fit1gs$null.deviance- fit1gs$deviance
D.df <- fit1gs$df.null - fit1gs$df.residual
p <- 1 - pchisq(D, df = D.df)
```