# WILL THEY COME?

# MODELING MATRICULATION DECISIONS FOR ADMITTED APPLICANTS AT THE UNIVERSITY OF ARIZONA

by

Omar Beltran

_____

A Thesis Submitted to the Faculty of the

DEPARTMENT OF AGRICULTURAL AND RESOURCE ECONOMICS

In Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

In the Graduate College

The University of Arizona

2017

## STATEMENT BY AUTHOR

This thesis has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona.

Brief quotations from this thesis are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permissions for extended quotation form or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgement the proposed use of the material is in the interest of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED:_____

## APPROVAL BY THESIS DIRECTOR

This Thesis has been approved on the dates shown below

_____ _____

Gary Thompson Date
Agriculture and Resource Economics

**ACKNOWLEDGEMENTS**

I would like to acknowledge that without the generous support of several people this thesis would not have been possible.

First, I would like to thank the phenomenal faculty and staff of the University of Arizona's Agriculture and Resource Economics department for supporting me through my undergraduate and graduate studies. In particular, I would like to thank the members of my committee: Dr. Gary Thompson, for his enthusiasm and diligence through every stage of the project. His high standards and attention to detail served as lessons I will take with me in my professional career; Dr. Satheesh Aradhyula, for always being approachable to us grad students and his talented way of sharing econometric knowledge in an easy-to-understand manner; Dr. George Frisvold, for always being willing to meet and providing valuable insights in both this project and previous research.

Second, I would like to extend my appreciation to Brian Berrellez, for not only providing the data used in this study but also consistently going out of his way to investigate and answer our questions. Without Brian's insights and efforts, this project would not have come to fruition. Also, Frank Santiago, for originally introducing the idea of a matriculation project and providing the recruitment data.

Finally, and most importantly, I would like to thank my mom for all her unconditional support throughout the years. It has not been an easy road to get to where I am, but thanks to her I was able to make it through.

**TABLE OF CONTENTS**

**LIST OF FIGURES**

**LIST OF TABLES**



**LIST OF APPENDICES**

## ABSTRACT

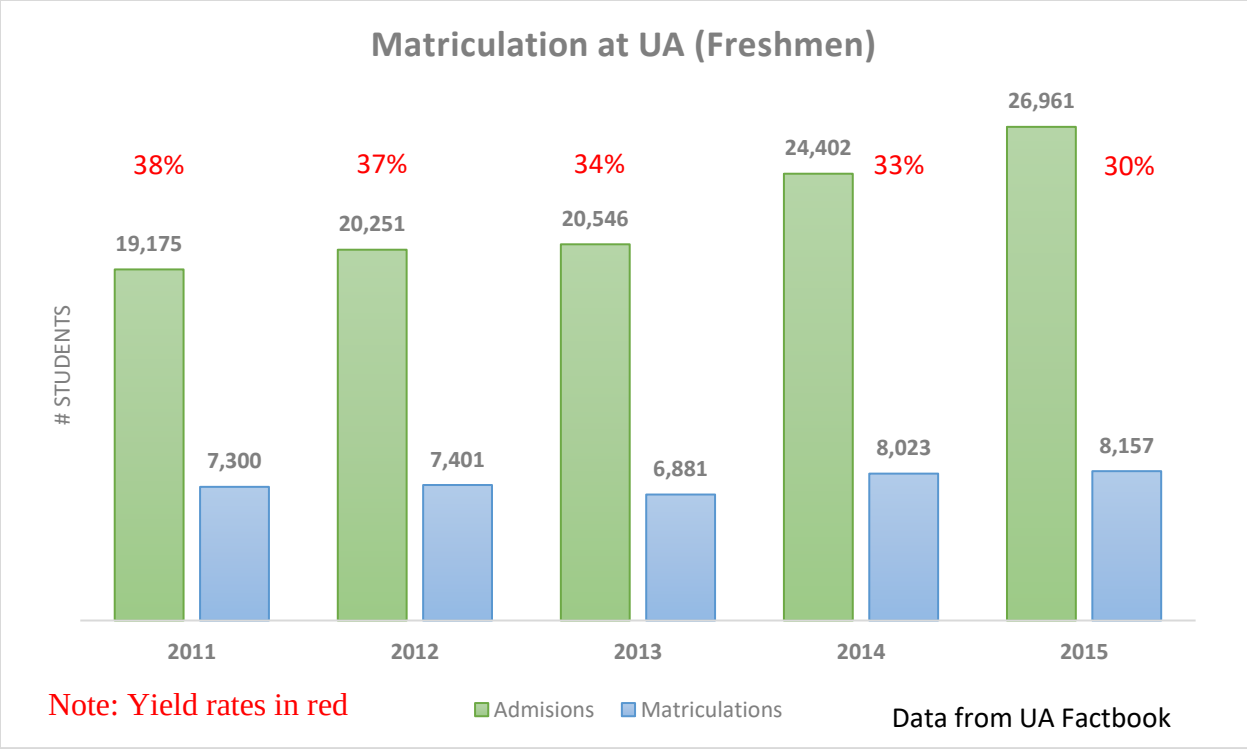
This study investigates factors influencing matriculation decisions for freshman applicants in the College of Agriculture and Life Sciences (CALS) at the University of Arizona. Two different modeling approaches are used on a five-year cross-sectional sample of applicants. Consistent with previous literature, a parametric logistic regression is specified to estimate the probability that a freshman applicant will matriculate in CALS. Additionally, this study also uses non-parametric gradient boosting methods to predict whether an applicant will matriculate. As a byproduct of using two different techniques to model matriculation decisions, an additional academic interest is to see how these two distinct approaches compare in terms of explanation and predictive capabilities. The results show that students who apply early and applicants with high standardized test scores are significantly less likely to matriculate. Moreover, applicants who attend campus tours, honor students, and students from high schools with many applicants are more likely to matriculate.

# Chapter 1: Introduction

## 1.1 Background

The University of Arizona (UA) is a large, public land-grant institution located in the Southwest region of the United States. UA currently enrolls over 40,000 students, both graduate and undergraduate. As of 2015, UA's acceptance rate for incoming freshmen was 76%(USnews College Rankings). In particular, the university enrolls between 7,000 and 8,000 new freshmen each fall. Figure 1 displays the admission and matriculation rates for freshman applicants in the past five years:

**Figure 1: Admission and Matriculation Rates from 2011-2015**




Despite admissions increasing significantly in recent years, yield rates (percentage of admitted students that matriculate) have fallen. A Rao-Scott likelihood ratio test was employed

to test for difference in proportions across the past five years. The results can be found in appendix 1, revealing a statistically significant difference in yield rates (1% significance level).

The decline in matriculation rates at the university could be due to many factors. For instance, the emergence of electronic common applications: "The college choice process is necessarily founded on the completion of applications. Where a student chooses to apply to college determines the set of schools in which he or she can choose to enroll" (Klasik, p.5). Electronic common applications allow students to apply to multiple universities at once, possibly indicating that an application to the UA may not signal as much interest or intent as it used to in the past. Furthermore, as the set of schools a student applies increases, the probability that a student selects any single school decreases.

Other potential explanations for the decline in matriculation rates could be the growth of alternative education options—such as community colleges and online program availability—and rising tuition costs relative to the other two public universities in the state. Tuition and fees for out-of-state students at UA increased 9% from 2011 to 2014, which is over twice as much as Arizona State University and nine times as much as in Northern Arizona University in the same time period (StartClass).

Furthermore, in recent years Grand Canyon University (GCU) has become a viable alternative for students seeking higher education. With several locations in Phoenix and one in Tucson, the private Christian university has done a lot of marketing to attract high quality students. GCU is more selective than UA (minimum GPA required is 3.0 compared to 2.0 for the UA) and offers hefty scholarships for students based on merit. GCU also has the same tuition price for both in-state or out-of-state applicants, which could be a reason why many applicants from the UA ultimately matriculate at GCU.

## 1.2. CALS College

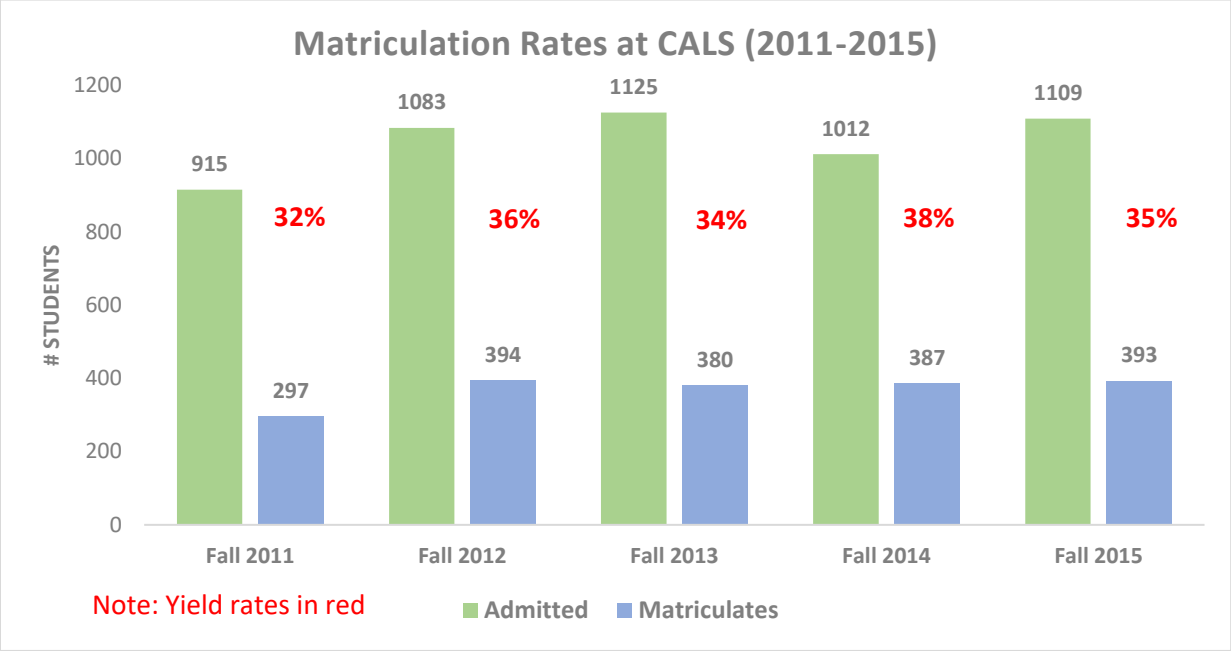
This study focuses specifically on applicants who selected a major in the College of Agriculture and Life Sciences (CALS). CALS is the 5th largest college at the University of Arizona, offering 14 different degree programs (see Table 1). Additionally, CALS provides diverse research opportunities and cooperative extension programs that help the local community.

**Table 1: List of undergraduate degree programs in CALS**

| | |
|---|---|
| Agribusiness Economics and Management | Microbiology |
| Agricultural Technology Management & Education | Natural Resources |
| Animal Sciences | Nutritional Sciences |
| Biosystems Engineering | Plant Sciences |
| Environmental & Water Resource Economics | Retailing and Consumer Sciences |
| Environmental Sciences | Sustainable Plant Systems |
| Family Studies and Human Development | Veterinary Sciences |

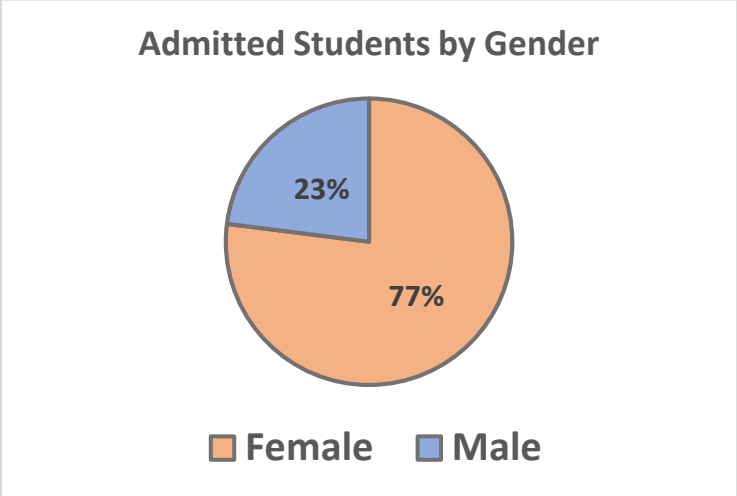
Figure 3 below shows admission and matriculation trends for CALS' freshmen in the past five years. The past four years particularly have seen a remarkably steady number of matriculates, with the 2015 yield rate being nearly identical to 2012. Contrary to aggregate university trends, a test of proportions reveals only a marginally significant change in matriculation rates (appendix 1, p-value<.067). The lack of statistical significance suggests that compared to the university as a whole, CALS appears to have more stable yields of incoming students.

**Figure 2: Matriculation Rates for CALS**




While the data used in this study will be described in more detail at a later chapter, it is important to note a couple of demographic variables in which the University and CALS differ systematically. Figure 4 indicates that over three quarters of CALS applicants are female. Perhaps this is not too surprising, given the largest programs in CALS—Animal Sciences, Veterinary Sciences, and Nutrition—might be more popular among female students. In contrast, according to the UA Factbook, the gender distribution at the University of Arizona is essentially balanced, with 52% applicants being female.

**Figure 3: Gender of CALS applicants 2011-2015**




Also from the UA factbook, a second demographic for which CALS and the University differ is in geographic region of origin. About 90% of those who matriculate in CALS come from the West region of the United States (Arizona, California, Montana, Washington, Oregon, Colorado, Utah, Nevada, Idaho, Wyoming) out of which 85% come from Arizona and California alone. While perhaps not a statistically significant difference, per the UA factbook, about 80% of undergraduate students at the University come from these two states[1].

A possible reason why the number of applications and matriculation rates in CALS do not appear to fluctuate much may very well be differences in demographics. As far as agricultural program availability, UA represents the best option compared with the other two public universities in the state. For out-of-state applicants, particularly from California, UA is less selective than its competitors—University of California Davis, Cal Poly San Luis Obispo and California State Fresno to name a few. The steady application and yield rates could also be

[1] Raw data for applicants at the entire university were not made available for this study. Therefore, no statistical test for differences between CALS and the University can be performed.

due to a consistent effort in recruiting or financial aid awarded.

### 1.3 Application Process

In order to adequately model matriculation for incoming students, it is imperative to have a thorough understanding of both the application and admission process. While the admission process keeps evolving and new nuances are introduced each year, the process was more or less uniform throughout 2011-2015, the time period for this study. Currently, almost all applications to the University are submitted online.

Requirements for admission at UA include four units each of English and math, three units of laboratory science and two units each of social sciences and a second-language. Standardized test scores are not required for admission. When applying, students self-report grades, high school rank and coursework. Moreover, applicants have the choice of disclosing their ethnicity and choose a major or remain undecided. Applicants must also indicate whether they wish to be considered for financial aid and if they intend to apply for federal aid. The university utilizes self-reported information for admission considerations, though an official high school transcript is required for enrollment.

While optional, if a student intends to apply to the Honors College –which provides advanced programs, smaller classes and research opportunities outside the classroom— standardized test scores—SAT or ACT— must be submitted. An additional requirement is that students complete a short writing assignment, which is typically a 500-word response to an argumentative prompt.

Figure 5 displays the fall application process for a typical high school senior. The UA's application for fall terms is open from late July to early May of the preceding academic year. In 2013, UA established *Wildcat Promise I* and *Wildcat Promise II* to encourage early applications.

Essentially, students who apply by certain deadlines in early October and early November receive priority and obtain their admission decision within three weeks. Students who submit applications at other times are typically reviewed afterwards, and will receive a decision from the UA within three to five weeks.

Upon admission, applicants also receive a notice of institutional merit financial aid. While not required for general admission (but required for the Honors college), students do need to submit standardized test scores for consideration of scholarships and merit awards. If applicants intend to apply for federal aid, they must fill out a *Free Application for Federal Student Aid (*FAFSA), which becomes available January[2]. Upon completion of the FAFSA, students will get notified about their federal-aid offers within two to four weeks, in the form of both the Pell grant and federal loans. The federal Pell grant program provides low-interest need-based grants to low-income undergraduate students (US Department of Education).

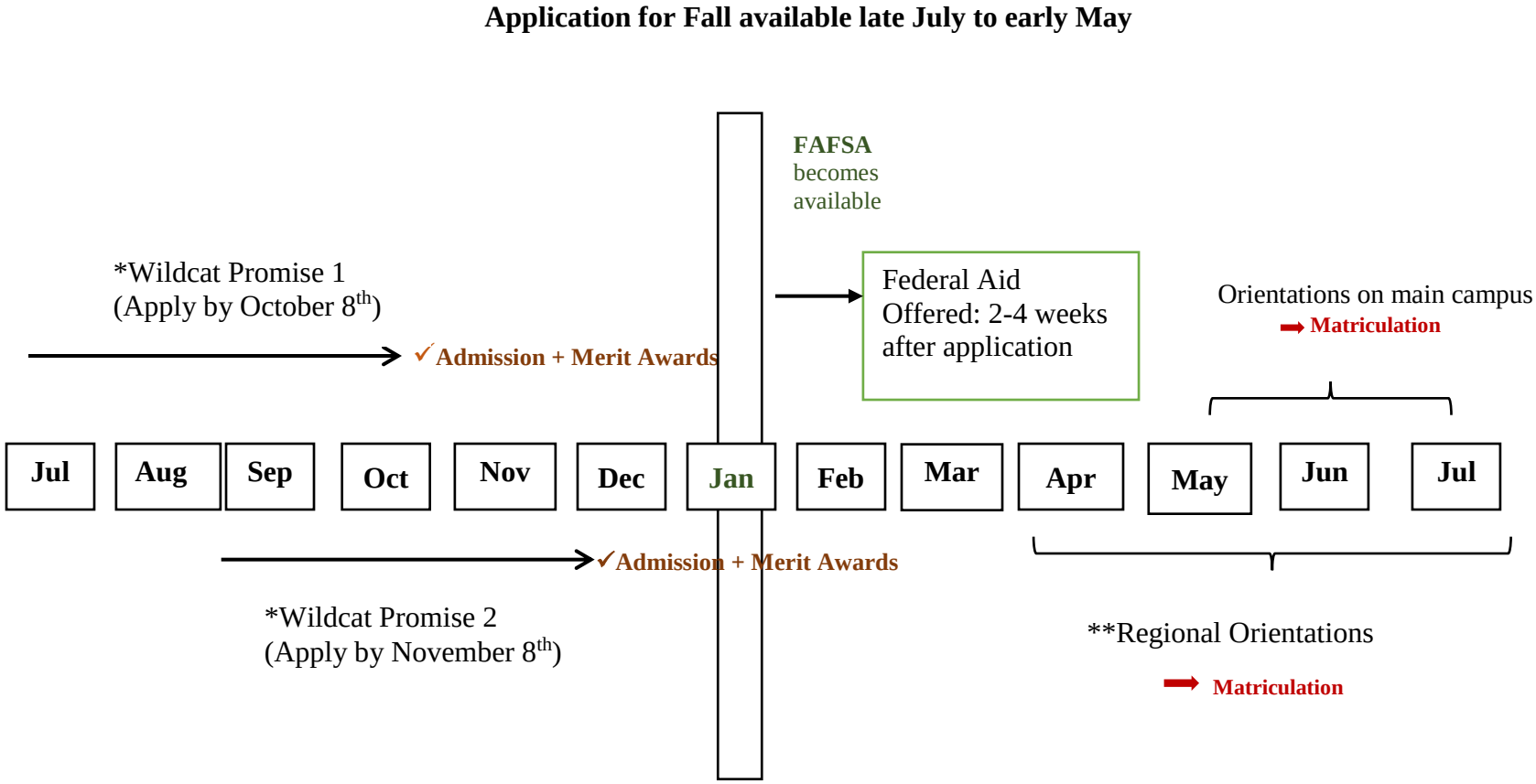
If a student intends to matriculate at the UA, he or she must complete a series of steps. To name a few, these include sending an official high school transcript, creating a university account, establishing residency (resident or non-resident) and paying an enrollment deposit. The final step involves signing up for and attending a New Student Orientation. Orientations are required for all incoming freshman and designate when matriculation officially takes place.

Orientation sessions are conducted all throughout summer. The New Student Orientation is an event where incoming students navigate around campus, learn about university policies, take placement exams, and ultimately sign up for classes. With that said, students can always add or drop courses at a later point. Orientations are typically conducted on site for most incoming

[2] Beginning in the academic year of 2017-2018, students can file a FAFSA as early as October. But for 2011-2016, the earliest date for submission was January 1st .

students. However, UA holds regional orientations for non-resident applicants during the spring and early summer. The regional orientations are similar to on-site orientations, with UA representatives and advisers traveling to four cities –Chicago, Seattle, New York City, Denver— and educating applicants about all university policies.

**Figure 4: Application Process**




* Established in 2013

**UA offers four regional orientations for out-of-state applicants in the spring and late summer. These take place in four cities: Chicago, Seattle, New York City and Denver.

To reiterate, from an administrative point of view, an incoming freshman student is designated as having matriculated once they attend the New Student Orientation. Administrators use orientation attendance numbers to forecast the size of the incoming class, which plays a crucial role in coordinating student programs and determining resource allocation. While it is possible that a student attend orientation and ultimately decide not to enroll, historically, 90-95% of students who participate in orientation do matriculate at UA. With that said, once classes begin in the fall, the admissions office will subsequently adjust enrollment numbers and obtain an accurate count of the current freshmen class. Specific counts are important for internal decision-making and need to be reported to federal agencies.

# Chapter 2: Literature Review

College matriculation studies first appeared in the late 1960's and early 1970's. Early research attempted to capture the factors determining if an applicant would pursue higher education. Subsequently, the research question shifted to where an applicant would attend. Explaining and predicting matriculation is nonetheless challenging, as there is much heterogeneity among freshman applicants and how they make decisions. Many factors influencing matriculation can be observed and quantified, though probably just as many are unseen and unobtainable by the researcher. Understanding past modeling approaches and results is essential, as it allows for improved models as well as comparisons with previous results. For the purposes of this study, an emphasis is placed on current literature and econometric models. The following provides a thematic review of the econometric literature regarding college matriculation decisions in recent years, as well as potential areas where this study contributes.

## Explanatory Variables in Econometric Models

In current literature, financial aid is without a doubt the most frequently examined variable. Not surprising, financial aid is often found to be a prominent predictor of matriculation decisions. College education is not cheap, and the price tag provides a constraint for many applicants. Additionally, how much aid to offer a student is one of the decisions over which colleges have complete autonomy. "Every year, thousands of high school seniors who have high college aptitude are faced with complicated arrays of scholarships and aid packages that are intended to influence their college choices." (Avery and Hoxby p.239). On the students' side, the less they or their parents pay out of pocket, the better off they are. Generally speaking,

studies find the more financial support students are offered and the lower their projected net cost of education, the more likely they are to matriculate (Van Der Klaauw; Avery and Hoxby; Weiler; Nurnberg et al; Desjardins).

Not all forms of financial aid are equal and can be divided into roughly two groups: free aid and non-free aid. The former takes shape in merit aid, grants and scholarships, those which are awarded to a student and do not have to be reimbursed. Conversely, non-free aid refers to aid which is offered to a student in exchange for immediate work or later repayment. The latter take form of loans —both private and federal — as well as work-study.

While more aid regardless of type has a positive impact on matriculation, as expected, free aid tends to have a stronger effect (Klaauw, Nurnberg et al, Monks). For instance, Monks studied the influence of merit aid at a small, selective university in the mid-Atlantic region. In an experimental setting, about 230 out of roughly 540 applicants were randomly selected for a $7,000 recognition award. The students who were selected for the award had a yield rate of 7.1%, compared to 3.2% for the non-aided group. The difference was found to be a statistically significant. Other researchers, however, have found no statistical differences between different types of aid. For example, in a private college setting, Linsenmeier et al, used a difference-in-difference approach to examine the influence of grants on enrollment. The authors found there was no statistically significant difference in enrollment for low-income minority students when the institution switched from offering loans to exclusively offering grants.

Relevant to financial aid and net cost, Curs and Singell estimated the price elasticity of tuition by modeling the application and enrollment decisions sequentially. However, in-state and out-of-state students were modeled separately. The authors argue that “in-state and out-of-state students are two significantly different student populations…” and “…combining these two

groups may bias the price elasticity towards zero" (Curs and Singell, p.112). Curs and Singell argue that out-of-state applicants are significantly more responsive to tuitions prices than in-state applicants and thus combining the two groups would understate the true price elasticity. The goal of the study was to analyze how applicants (1995-2000 academic years) responded to changes in price at the University of Oregon, as well as tuition rates at competing institutions. For in-state students, the net price at the University of Oregon —defined as tuition and fees minus financial aid awarded — and the average tuition and fees at the two other public universities in the state were included. Similarly, for out-of-state students, the average tuition and fees of the top 20 institutions that share a common pool of applicants with the University of Oregon were included in the model.

While Curs and Singell found the average price of competing institutions was significant in the application decision, it was not significant in determining enrollment. The authors discovered out-of-state applicants were relatively more sensitive to price changes than in-state applicants. Response to price increases at the University of Oregon was found to be inelastic in the enrollment model, with a 1% increase in net price resulting in a 0.23% decrease in enrollment for in-state students and a 0.62% decrease in enrollment for out-of-state students. A potential source of measurement error of this study could be the lack of variability in the average price of competing institutions. Given that financial aid offers from other schools was unobserved, every student in their sample who applied to a competing institution was assigned the same net cost. Without complete information about students' alternative financial aid packages, it is difficult to estimate or compare the true net cost differences across universities.

Not only have different types of aid been studied, but also the way financial aid is packaged and delivered. In a pioneering study in 2002, Avery and Hoxby comprehensively

examined the impact of financial aid on enrollment decisions. The authors designed surveys for high-ability seniors applying to college in 1999-2000. They worked with over 500 high schools nationally with a reputation of sending students to selective colleges. Counselors from these high schools randomly selected seniors in the top percentage of their classes and administered three questionnaires throughout their last year of high school. The first questionnaire was administered January and asked for family background, academic, and extracurricular information, along with questions about preferred colleges and schools they had applied to. The second survey was administered in May and asked about student's outcomes, financial aid awards from all schools, admission and matriculation decisions. Also, the survey asked how much financial aid played a role in their decision. A third questionnaire was administered to each parent about their child's choice as well as income information. All in all, 3,240 responses from 396 high schools were collected and usable for analysis.

Avery and Hoxby's extensive survey work produced several interesting findings. As mentioned, after controlling for other factors, more grants and loans offered increased the probability of matriculation. An additional thousand dollars in grants and scholarships raised the probability of matriculation by 11% while and additional thousand dollars in loans raised probability of matriculation by 7%, though the difference was not statistically significant. Based on responses from parents, the authors found students whose parents attended low selectivity colleges were more responsive to grants and loans. In other words, financial aid for this group had a greater effect than those with parents who attended highly-selective colleges. Avery and Hoxby also discovered students had different perceptions and reacted differently to the way net cost was presented. For instance, an additional $2,000 grant was found to have a stronger effect than a $2,000 reduction in tuition. Moreover, they found students responded more to named

scholarships, and naming a grant increased the predicted probability of enrollment by 86%. They also found that front-loaded grants, in which most of the aid is delivered upfront, also increased the predicted probability of enrollment by 48%.

In addition to financial aid, high school characteristics and historical connection have also been studied. As Wolniak and Engberg describe, "Quite simply, it appears that the more established a high school's historical connection with a particular college, the more likely a student from that high school will choose to enroll"(Wolniak and Engberg p.43). Particularly, Wolniak and Engberg studied the effects of *feeder networks,* or high schools that "feed" or provide many applications, admissions, and ultimately enrollments to universities. Using application data from 8 private universities, the authors constructed an index measuring historical connection in the past 5 years with these private universities. They found the set of schools students applied to was indeed influenced by what peers and students in previous years had done. "A student's choice set of colleges and universities is affected by the share of peers in her or his graduating class who signal interests in the same institutions"(Wolniak and Engberg p.31). The feeder-network effect was found to have a positive impact on enrollment, and applicants from high schools with higher indexes were more likely to enroll at the universities in the study. The authors controlled for the confounding effects of student's background characteristics and the feeder effect remained significant even after controlling for college-specific characteristics (Wolniak and Engberg).

Similarly, Irina Johnson analyzed the influence of high school-specific characteristics on enrollment and retention at a non-selective private university. Using a five year sample of about 15,000 students from 400 different in-state high schools from 2001-2005, Johnson found applicants from high schools where a high percentage of students took the SAT examinations

were more likely to matriculate, though the effect was marginally significant. SAT scores were found to have a concave effect: As a student's test scores increased, so did the probability of enrollment, but if scores exceeded a certain threshold, the probability of enrollment declined. Distance from campus was also found to be significant: students from high schools within 60 miles of the university were 81% more likely to enroll. Though percent of peers receiving free lunch at an applicant's high school was examined, this was found not to be a significant predictor of enrollment. Relevant to income, Johnson also found a statistically significant concave relationship between parents' income and enrollment, though the magnitude of the squared income term was small. Additionally, applicants whose income was higher than average peer income were slightly more likely to matriculate and a lot more likely to matriculate than those with less than average incomes.

The influence of a university's ranking on matriculation has also been studied. Griffith and Rask studied the importance and influence of the U.S News and World Report (USNWR) rankings on matriculation. The authors note that every year "high-ability high school seniors also await the arrival of this annual issue because it serves as a guide to the schools they are considering for college" and "applicant pools change when a school drops in ranking" (Griffith &Rask p. 2). These rankings are very thorough and have multiple categories, including student-to-faculty ratio, average standardized test scores and diversity ratios, to name a few. Many universities allocate resources to retain their status, improve, or in some cases recover their ranking (Griffith & Rask).

Griffith and Rask modeled matriculation for around 8,000 admitted students at Colgate University in 1995-2004 . The authors used responses from the Admitted Student Questionnaire (ASQ)

and divided the sample into two categories; full-pay — students who did not apply for financial aid — and need-based — students who applied for financial aid. The authors found the influence of the USNWR on enrollment to be robustly significant. The impact of changes in rank was different for the two groups. Particularly, students in the full- pay sample were more sensitive to rankings than the need-based. Full-pay students experienced significant reductions in probability of matriculation for every drop in rank in the top 20 schools. However, for this group, schools' changes in rankings outside the top 20 appeared to have no significant influence on enrollment. Other findings from this study revealed women were less sensitive to rank than men, minorities were found to be sensitive to changes in the current minority population, and universities' rank has become more important over time from 1995 to 2004.

The influence of recruiting on matriculation has briefly been explored. Using data from Fall 2003, Goenner and Pauls studied whether recruitment activities such as inquiries and campus visits were indicators of enrollment. At the University of North Dakota, the authors used information from students who either inquired about the university or attended a campus visit to specify several enrollment models. The data on home addresses provided from these inquiries was sparse and limited, so the authors used zip code demographic variables as proxies for individual demographics for about15,000 students who contacted the university.

Goenner and Pauls randomly divided their sample in half. One half of the sample was used for obtaining a model and the other half was used for testing its predictive power. The authors found that the number of inquiry contacts between the student and the university, as well as the number of campus visits remained significant in several model specifications. An interaction term between distance from campus and a campus visit was highly significant. Unfortunately, the Goenner and Pauls did not quantify how much of an influence these

recruitment measures had on matriculation. However, their model correctly predicted enrollment 89% of the time in the test sample.

There appears to be one study that has used machine learning techniques in modeling matriculation, but not directly. In *Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School,* San Pedro et al. discussed the effects of a New England mathematics middle school tutoring system (ASSISTment) on college enrollment. "ASSISTment" is an interactive online system that poses questions to students and guides them through math problems, providing hints and suggestions along the way. The authors argue that based on patterns of engagement and inactivity at this early age, one can predict whether a student will enroll in college:

> "Many factors influence a student's decision to enroll in college. A lot of them external or social factors: financial reasons, parental support and school support. Another major factor, however, is one's ability and engagement, which develop over early years, and begins to manifest strongly during the middle school years" (San Pedro, et al, p. 6)

Analyzing close to half a million logs from the tutoring system, the authors use machine learning methods to construct measures of engagement, carelessness, boredom and confusion. The measures of activity combined with enrollment records from the National Student Clearing House database were then incorporated into a logistic regression to model enrollment. The authors found that boredom and disengagement in middle school were associated with lower probability of college enrollment.

An important remark to make is that within all the previously mentioned studies, other control variables have been examined. Demographics, minority status, early applicants, whether

an applicant has parents or siblings who attended college, distance from campus and ability in terms of standardized test scores have been found to influence matriculation. While the effects on matriculation are nonetheless interesting, these variables are mainly included as control measures, though they can play a role in the selection of a model.

This study builds upon the existing literature, and conflates many concepts for a comprehensive study of matriculation at the University of Arizona (UA), a large public university. This study is unique in that it exclusively examines applicants with a specific set of intended majors. The availability of data provides an excellent opportunity of segmentation. The underlying assumption is that students who apply for related majors are likely to have similar interests and aspirations, and are probably similar in other unaccounted ways as well. While the sample used in this study is not a random sample, by focusing on a specific group of like-minded applicants one can reduce the heterogeneity across applicants and their decisions.

Rather than specializing on a specific facet of matriculation decisions, this study comprehensively analyzes many of the most commonly influential factors on matriculation found on the literature. Not many prior studies have conducted a thorough, inclusive analysis of matriculation at a large public university. This study also quantifies the impact of recruitment measures like campus tours. Other studies relevant to recruitment have ambiguously suggested campus visits and inquiries are indicators of enrollment intentions, but did not quantify the impact of recruiting on matriculation.

This study is unique in that it uses machine learning techniques as well as parametric models to explain and predict enrollment. Machine learning has been found to be a strong prediction tool in many industries, and an area that academic research on matriculation up to this point has under-utilized. While the more conventional binary response model is also used in this

study, the addition of machine learning methods provides additional insights and adds to robustness of results.

# Chapter 3: Data and Econometric Methods

The purpose of this study is to determine and quantify factors influencing matriculation in the College of Agriculture and Life Sciences (CALS) for incoming freshman. A secondary objective is to construct a model that can accurately predict out of sample. Given that only CALS applicants are observed, results from this analysis may not be applicable to the University of Arizona as a whole. Nonetheless, findings from this study can be useful and ultimately lead to actionable suggestions that can assist enrollment managers and CALS recruiting.

Two different modeling approaches are employed. First, consistent with many matriculation studies, a logistic regression model is specified to explain and predict whether an applicant matriculates at UA. Second, and not common in previous literature, this study employs machine learning methods to reinforce the analysis. In particular, gradient boosting models are used to predict matriculation decisions.

A logistic regression approach can be referred to as parametric, which makes assumptions about the probability distribution from which the data were drawn (Hoskin). Conversely, a machine learning approach is non-parametric, making no underlying assumptions or restrictions. With machine learning, an emphasis is put on pattern recognition and prediction. As a byproduct of using two different techniques to model matriculation decisions, an additional academic interest is to see how these two distinct approaches compare in terms of explanation and predictive capabilities.

## 3.1 Data and Sample

Data for this study was provided by the data analytics department in CALS. The sample selected is a five-year cross-sectional dataset restricted to a) 2011-2015 fall admitted applicants, b) students who selected a CALS major, and c) domestic freshmen. Data access and availability played a major role in the selection of this sample. With that said, the methodology was to select applicants that experience a homogeneous admission process, which is crucial to discern true factors affecting matriculation. Ultimately, the goal is to be able to construct a model that can predict matriculation in subsequent years and aid future internal decision-making within the college.

The sample period selected was started in 2011 because of a change from previous years in the way the University of Arizona (UA) tracked and collected student-specific data. In 2011, UA implemented *UAccess* which is the is the "primary method of delivering quality financial, employee, student, and research data to colleges and departments across campus" (University Information Technology Services). The application and admission process from 2011 to 2015 remained mostly unchanged.

As mentioned in the previous chapter, when applying to the UA, students have an option of selecting a major or remain undecided. Data for this study was only available for applicants who specifically selected a CALS major. It is important to mention that this sample is not random, as students self-select by choosing a major in CALS. In addition, this sample is a small subset of all applicants to the university and therefore, findings from this study may not represent matriculation tendencies of all freshman applicants. If data becomes available, future research should compare factors affecting matriculation decisions across all the different colleges at UA.

Given the heterogeneity in how applicants make decisions, it is important to select a group that has similar observable characteristics. While international and transfer students make up a significant portion of students—10% and 20% respectively— and are an asset to the student body, these applicants are systematically different than typical high school seniors. As Unda explains in a prior case study of enrollment at UA, "The reasons to exclude these important cohorts of applicants is due to the differences on admissions requirements and processing" (Unda, p.19). For example, transfer students are usually older and do not send standardized test scores. International students often have additional language requirements and must go through specific procedures establishing residency. Moreover, it is also fair to say that most recruiting, at least in CALS, is targeted to high school seniors. Having students in a sample that were not targeted might confound the impacts of recruiting on matriculation.

## 3.2 Variables

Variable selection in this study was based on both availability and current literature on matriculation. For purposes of prediction, it was important to include only variables measured before the matriculation decision. The dependent variable is a dichotomous variable indicating whether the student matriculated at UA or not. Table 2 on page 36 provides a comprehensive list of all the variables used in the analysis, though different combinations of variables were used in several model specifications. Additionally, below is a description and formulation of several key variables.

*Standardized test scores*

Though officially not required for admissions, either the SAT or ACT (or both) composite test scores were submitted by most applicants. For the purpose of having a uniform ability

measure, SAT scores were converted to ACT scores. Given that the same ACT score corresponds to multiple SAT ranges, it was not clear how to map ACT scores to SAT scores. If an applicant had both an SAT and ACT, the SAT score was converted to an ACT score, and the maximum score out of the two was selected. After conversion, 82% of applicants had a standardized test score ranging from an ACT score of 7 to a perfect score of 36. A full conversion table can be found in appendix 2.

*Financial Aid Variables/Income*

After careful examination, it appears financial aid information in the data used in this study was truncated, and financial aid offers—both institutional and federal—were observed almost exclusively for students who matriculated. Only 26 of 3,349 applicants who did not matriculate had financial aid data. In contrast, 80% of applicants who matriculated had financial aid data. Incomplete financial aid data would bias the analysis because using a truncated sample would make it seem as if an applicant received financial aid he or she would almost certainly matriculate. Moreover, a model conducted using this data would most likely overestimate the impact of financial aid, while possibly undermining the effect of other variables in the model. While financial aid is crucial for a comprehensive study of matriculation, one needs complete aid information for both those who matriculated and those who did not. The university ought to keep all records of financial aid, even if a student does not matriculate, to be able to analyze the impact of financial aid on matriculation decisions.

Since family income is not provided, median household income by zip code –obtained from the 2010 American Community Survey (ACS)— is used as a proxy. The median household income is time invariant because of lack of available annual data, and applicants from the same zipcode were assigned the same household income irrespective of application year.

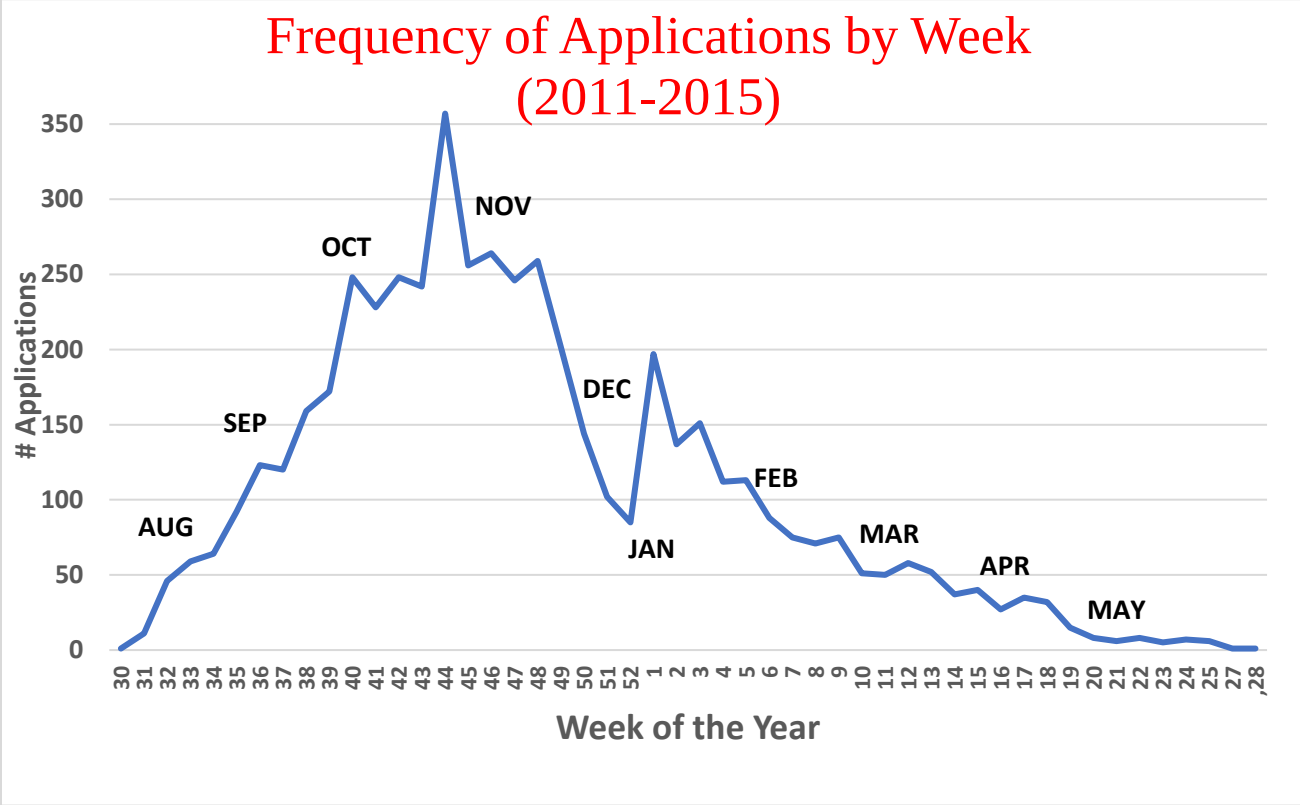
*Honors and High-Ability Students*

To distinguish between the academic quality of applicants, a variable is used for those who applied and were admitted to the Honors College. Likewise, high ability students are defined by those students who received outstanding national and state awards based on merit, standardized test scores and impact on their communities. The high ability student indicator includes Baird nominees and Baird scholars. Both recognitions are given to outstanding students who graduate in the state of Arizona. Similarly, the high ability student indicator includes Flinn finalists and Flinn scholars, awarded by the Flinn Foundation which provides generous scholarships to selected recipients. Also included in this group are National Merit finalists and semifinalists.

*Timing of Applications*

How early or how late a student applies to the university may signal a level of interest or intent to matriculate. Figure 6 below shows the timing of applications submitted over the past five years. In CALS, over a third of applications are submitted in October and November. Following November there is a decline in applications over winter break and the holidays. Shortly after, there is a significant spike January first—perhaps students waiting for the FAFSA to become available to apply— followed by a steady decline towards May, with slight peaks, most likely for last-minute decisions to go to college.

**Figure 5 : Applications by Week (2011-2015)**




The application date for each student was used to measure the number of days from when an application was officially submitted to the first day of classes of the intended fall term. Additionally, given that the timing of the application might signal different intentions depending on the type of students, an interaction between ACT score and days ahead is examined. A hypothesis is that students who score higher on standardized test scores and apply early might matriculate less often. These students are more strategic and well prepared, and likely have more schools in their choice set. Because of their high-test scores, these applicants have a higher probability of being accepted at other highly ranked colleges so they are expected to be less likely to matriculate at UA.

*Recruitment Variables*

Two measures of recruitment were examined in this analysis: campus tours and Arizona Experience (AZX). Campus tours are opened to the public and are offered twice a day, six times a week throughout the year. Campus tours are roughly two-hour events and are free to attend. At the beginning of the tour, UA student ambassadors make a short presentation to prospective students and their guests, followed by a tour around the main campus. Those who attend a campus tour are informed about general UA policies, history, and traditions.

By contrast, AZX visits are comprehensive all-day events, more analogous to an open-house experience. AZX visits cost $20, as opposed to the free campus tours. Students and guests have an opportunity to attend a myriad of presentations and information sessions. Representatives from each college at UA hold sessions in the morning and late afternoon, where students can learn about individual programs and academic opportunities at the university. Also, students can visit dorms, the recreation center at the UA, talk to advisors from each college, get help on their applications and even have a courtesy meal at the student union.

Recruitment workers call late February to mid-April “Paratour Season”, which is when most recruiting visits, both campus tours and AZX, take place. During this period, the UA gets anywhere between 400-600 attendees per week. Given limited availability of data, campus visits and AZX tours in were only tracked from 2013-2015. Thus, a separate analysis was used to examine the impact of recruiting on matriculation for these cohorts.

*High School Peer Effect*

Previous literature has found high school variables have an impact on matriculation decisions. For example, Wolniak & Enrgberg showed that historical connection with a university

had a positive impact on enrollment. The literature, however, has not examined a peer effect; that is, how does someone in your high school going to a particular college affect your decision to apply or even matriculate at that same college. Furthermore, having a measure of a number students from the same high school might shed light into other factors affecting matriculation not observed. For example, a high school counselor that is a supporter of the UA and encourages students to apply may increase the likelihood of matriculation. Also, having friends or acquaintances from the same high school who apply and matriculate at UA may influence an applicant to matriculate.

This study employed a count variable *hs_peers* which measures how many applicants were admitted from the same high school in the same year, minus one. For instance, if three applicants were admitted to UA in 2012 from High School X, the *hs_peers* variable would be 2 for each applicant, meaning that they each had two peers applying to UA. Similarly, *Past_Peers* keeps a count of students from the same high school but that were admitted to UA in prior years. The lag for this variable ranges from one to four years, for cohorts 2012 and 2015 respectively. For instance, if a 2015 applicant from high school X had two former peers in 2014 that were admitted to UA and 3 former peers in 2013, the past peers would be 5 for that applicant. Past peers measures how much influence on matriculation decisions, if any, is exerted by admission decisions of students in years prior.

Given the abundance of incomplete high school names, misspellings, and even the same high school name in different states, special attention and care was given when assigning students to a particular high school. High school names were carefully examined to make sure spelling and abbreviations represented the same high school. Additionally, high school names were cross-checked by home state to make sure they represented the same school. Data on high

school size and type of high school—public, private, catholic, charter, magnet—was obtained from National Center of Education and Statistics (NCES).

Other control variables include geographic region, gender, ethnicity, major, and an indicator whether the student is the first in their family to go to college. Year dummies are also included to control for potential year effects.

**Table 2: List of Variables**

| Variable | Description |
|---|---|
| Matriculation | Dependent Variable. Dummy: 1 if matriculated, 0 otherwise |
| ***Demographics*** | |
| Northeast | Dummy: 1 if home state in ME, CT, MA, NH, VT, RI, NJ, NY, PA |
| Midwest | Dummy: 1 if home state in IL, IN, MI, OH, WI, IA, AK, KS, MN, MO, NE, ND, SD |
| South | Dummy: 1 if home state in DE, GA, FL, MD, NC, SC, VA, WV, MS, KY, AL, TN, AR, LA, OK, TX |
| West | Dummy: 1 if home state in AZ, CO, ID, MT, NV, NM, UT, WY, CA, OR, WA (reference group) |
| White | Dummy: 1 if Caucasian, 0 otherwise (reference group) |
| Black | Dummy: 1 if African American, 0 otherwise |
| Hispanic_Mexican | Dummy: 1 if Hispanic/Mexican, 0 otherwise |
| Asian | Dummy: 1 if Asian American, 0 otherwise |
| Other_Ethn | Dummy: 1 if other ethnicity or missing, 0 otherwise |
| Female | Dummy: 1 if female, 0 otherwise |
| d_First_Generation | Dummy: 1 if applicant is first one in family to go to college |
| Median_Household_Income | Median household income by applicant's home zip code |
| ***Ability*** | |
| d_High_Ability_Student | Dummy: 1 if applicant is Baird nominee, Baird scholar, Flinn finalist, Flinn scholar, national merit finalist or national merit semifinalist. |
| d_Honors_Admit | Dummy: 1 if applicant was admitted to the Honors College |
| ACT_Max | ACT composite score (SAT scores were converted to ACT. If applicant had both maximum score was selected) |
| ACT_sq | ACT_Max squared |
| AP_Units | Number of AP credits from high school |
| ***Majors*** | |
| d_ABEM | Dummy: 1 if selected major in Agribusiness Economics and |

| | |
|---|---|
| | Management |
| d_AGTE | Dummy: 1 if selected major in Agricultural Technology Management & Education |
| d_ASC | Dummy: 1 if selected major in Animal Sciences |
| d_ENV | Dummy: 1 if selected major in Environmental Sciences |
| d_MICR | Dummy: 1 if selected major in Microbiology |
| d_NTR | Dummy: 1 if selected major in Natural Resources |
| d_NUSC | Dummy: 1 if selected major in Nutritional Sciences |
| d_PLS | Dummy: 1 if selected major in Plant Sciences |
| d_PRFS | Dummy: 1 if selected major in Family Studies |
| d_PRRC | Dummy: 1 if selected major in Retailing and Consumer Sciences |
| d_VSC | Dummy: 1 if selected major in Veterinary Sciences |
| d_Other_Major | Dummy: 1 if other major |
| ***Year Dummies*** | |
| d_2011 | Dummy: 1 if student applied for Fall 2011 |
| d_2012 | Dummy: 1 if student applied for Fall 2012 |
| d_2013 | Dummy: 1 if student applied for Fall 2013 |
| d_2014 | Dummy: 1 if student applied for Fall 2014 |
| d_2015 | Dummy: 1 if student applied for Fall 2015 (reference group) |
| ***Timing of Application*** | |
| Days_ahead_app | Count variable: Days between application submission and first day of classes |
| Days_app_Sq | Square of Days_ahead_app |
| Interaction_Act_Days_App | Days ahead*ACT_max score |
| ***High School*** | |
| HS_Size | Number of students in high school |
| d_Public_school | Dummy: 1 if applicant graduated from a public school (reference) |
| d_Private_school | Dummy: 1 if applicant graduated from a private school |
| d_Catholic | Dummy: 1 if applicant graduated from a Catholic school |
| d_Charter | Dummy: 1 if applicant graduated from a charter school |
| d_Magnet | Dummy: 1 if applicant graduated from a magnet school |
| HS_Peers | Count variable: Number of high school peers who were admitted to the UA in the same year |
| Past_Peers | Count variable: Number of high school peers who were admitted to the UA in previous years |
| ***Recruitment*** | |
| Campus_Tour | Dummy: 1 if applicant attended a campus tour |
| AZX_Visit | Dummy: 1 if applicant attended Arizona Experience (AZX) |

### 3.3. Parametric Approach: Logistic Model

Modeling matriculation decisions is particularly challenging due to the fact of asymmetric information. More often than not, university data is asymmetric in the sense that it only includes student-provided application responses. For example, while universities know internal financial-aid packages, they lack awareness about other schools' financial aid packages in an applicant's choice set. Moreover, the student himself is the only one who truly knows his preferences and intentions. Incomplete information is a serious hurdle to overcome because factors that might affect whether a student matriculates at a specific school are often unobserved in the data.

Despite asymmetric information between the applicant and the institution, many matriculation studies using only internal institutional data are able to provide insights into matriculation decisions. Following previous studies, the first part of this analysis is conducted using a logistic model. Logistic models allow us to distinguish among which factors have a significant impact on the decision to matriculate. Moreover, logistic models can be used to quantify the effects of the independent variables on matriculation and assign probabilities of enrollment to students with specific attributes. As Desjardins describes "This analytic approach can also be used to predict each student's probability of enrollment, thereby allowing us to understand better the enrollment propensities of different groups of students" (Desjardins, 2002 p.538).

Logistic models are simply regressions where the dependent variable is dichotomous, either zero or one; in this context, whether an applicant matriculates or not. The dichotomous variable is subsequently regressed on a set of explanatory variables whose effect can be interpreted as changing the probability of the outcome. The probit model is another type of

binary response model also common in the literature. The main difference between logistic and probit models is the underlying assumption regarding the distribution of the error term. Logistic models assume the error term follows a logistic distribution, whereas probit models assume a standard normal distribution.

Once the model has been specified, maximum likelihood methods are used to obtain parameter estimates. While the sign of the estimated coefficients indicates the direction of influence, the magnitude of the coefficients cannot be interpreted directly. Therefore, a useful additional step is to compute marginal effects, which are changes in probability of the dependent variable due to changes in the independent variable, holding other things constant.

### 3.3.1 Model Specification

The first step in the parametric portion of the analysis is to define a logistic model. Consider the linear function below, which shows matriculation as a linear combination of observed covariates $\boldsymbol{x}'_i$ and **β,** a vector of regression coefficients and a corresponding error term $U_i$.

$$Matriculation_i = \boldsymbol{x}'_i\boldsymbol{\beta} + U_i$$

However, since matriculation is specified as linear function, the above equation can lead to predicted probabilities below zero or above one, which does not make sense in the context of estimating how likely an applicant is to matriculate. To be consistent with the axioms of probability, instead consider a function *p,* as the ratio of an exponential function of covariates and the vector of regression coefficients, over the same quantity plus one:

$$p_i = \frac{e^{x'_i\beta}}{e^{x'_i\beta} + 1}$$

The function *p* is always between 0 and 1, satisfying the axioms of probability. For a logistic model, the next step involves defining and converting the dependent variable—in this case,

matriculation—to odds. Consider the function below, which can be interpreted as the odds of matriculation.

$$Odds_i = \frac{p_i}{1-p_i}$$

For example, if the probability of matriculation for an applicant is 50%, the odds of that applicant matriculating would be 1 to 1. If the odds of matriculating are 25%, the odds reduce to 1 to 3. Algebraically, it can be shown that the odds function is equal to the exponential function of covariates and vector of regression coefficients

$$\frac{p_i}{1-p_i} = e^{\boldsymbol{x'}_i\boldsymbol{\beta}}$$

Taking the natural log of both sides yields our desired logistic model, in which the log of odds is regressed linearly on the set of covariates. The log of odds is often referred to as a latent variable, which is not observed, but is directly related to the true variable matriculation.

$$Matriculation^* = \ln\left(\frac{p_i}{1-p_i}\right) = x'_i\boldsymbol{\beta} + U_i$$

Now defining $\boldsymbol{x'}_i\boldsymbol{\beta}$ specifically in this study:

$$\boldsymbol{x'}_i\boldsymbol{\beta} = \alpha + \boldsymbol{\beta}_i[Demographics] + \boldsymbol{\gamma_i}[Financial\ Aid] + \boldsymbol{\varpi_i}[Ability] + \boldsymbol{\eta_i}[Recruitment] + \boldsymbol{\phi_i}[High\ School\ Variables]$$

**α, β, γ, η, ϖ, ϕ** are vectors of estimated coefficients. *Demographics* include student-specific demographic variables. *Financial aid* includes financial aid offered and median household income by zip code. *Ability* contains ACT score, number of AP units, and honors/high-ability indicators. *Recruitment* includes campus tour and AZX visits while *High School Variables* include high school size, type and peer and past peers variables. SAS University Edition and SAS 9.4 were used to estimate coefficient parameters. Stata14 was used to compute marginal

effects from this regression.

### 3.3.2 Analytical Strategy

Traditionally in matriculation studies, a logistic model is run on a historical applicant dataset and the whole sample is used in obtaining parameter estimates. Subsequently, marginal effects are obtained to describe changes in probability of matriculation given changes in explanatory variables. While this methodology might be sufficient for explanation purposes, it may be deficient if one wants to predict out of sample. A potential issue with utilizing the whole sample for estimation was outlined by Desjardins; "When you use the same data to test predictive accuracy of your model that you use to fit the model, it biases your results" (Desjardins, p.539). In other words, a model will naturally do a good job predicting its own data, often overfitting. If the model is too specific to its unique sample structure, is difficult to use the estimated coefficients to predict in other samples.

As an improvement to prior studies, and borrowing extensively from Desjardins's recommendations, the sample in this study is partitioned. 2011-2015 data is randomly divided into two segments: a *training set*, which is the set used to estimate parameter coefficients; and a *validation* set, used to test how well the training model classifies those who enroll and those who do not enroll. It is important to partition the sample to provide valid assessments of the performance of the predictive models (Steinberg). Seventy percent of the sample was randomly selected for training, while thirty percent was held out for validation to predict out of sample. The validation set was "scored" which essentially means applying estimated coefficients from the training sample to the validation data set to compute predicted probabilities (SAS Users Guide).

To measure the efficacy and fit of the model, three different standards are used. Again, based on Desjardin's work, the first measure is a classification table. This is a table that keeps track of right and wrong predictions. Using parameter estimates from the training set, one can assign probabilities of matriculation to applicants in the validation set. Then, a cutoff threshold between 0 and 1 is selected. If the predicted probability of an applicant exceeds the selected threshold, the applicant is predicted as having matriculated. On the other hand, if the predicted probability is less than the threshold, the applicant is predicted as not matriculating. After classifying each applicant, the table keeps track of how many times the model predicted correctly based on the predicted probability, as well as incorrect predictions.

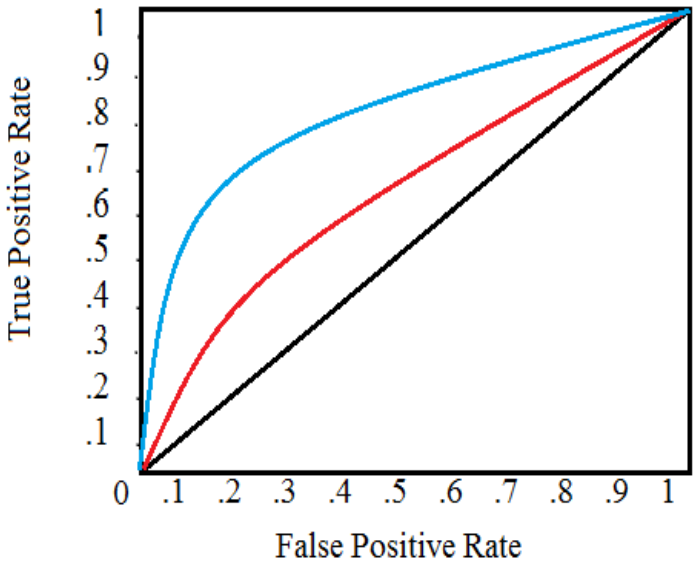
Classification tables are sensitive to cutoff values. A low cutoff value is conservative and will guarantee that most students who enroll are not misclassified. However, it will also predict many applicants enroll who actually do not (false positive). Conversely, having a large cutoff value will reduce the number of false positives. But at the same time, a large cutoff value is strict and will classify many applicants who enroll (but have predicted probabilities on the margin) as not enrolling. The choice of cutoff threshold depends on goals and costs of misclassifying enrollees and non-enrollees (Desjardins, 2002).

Given the sensitivity of predictions to cutoff values, three cutoff values are used: 0.35, 0.5 and 0.65. Though 0.5 is the most commonly selected cutoff value, it may not be appropriate if the training sample contains an unequal number of enrollees and non-enrollees (Desjardins, 2002). A cutoff of 0.35 is also used, as it is the historical yield rate of matriculation in CALS over the sample period. This represents a relatively "low" cutoff value. Lastly, 0.65 is selected as well to have a relatively high cutoff value.

A secondary way to test the fit of the model are receiving operating characteristics curves

(ROC). "A receiving operating characteristics (ROC ) graph is a technique for visualizing, organizing and selecting classifiers based on their performance" (Fawcett, p. 862). ROC curves are related to classification tables. They are two-dimensional graphs in which the true positive rate –applicants that the model predicted enrolled and did— is plotted on the y axis and the false positive –non-enrollees predicted as enrolling— is plotted on the x axis. Figure 6 displays an example of multiple ROC curves.

**Figure 6: ROC example**




There are three important curves to point out in the figure, each corresponding to a different model. First is the y=x diagonal axis in black. This line is effectively equivalent to flipping a coin to predict the outcome or randomly guessing. The model predicts half of true positives correctly, but also predicts false negatives half the time.

The red curve above is an improvement to randomly guessing, depicted by the fact that

true positives occur at a higher frequency than false positives. Similarly, the blue curve is more accurate than both the red and black curves. Ideally, an optimal ROC curve would be as close as possible to the y-axis, picking up true positives most the time while simultaneously minimizing the false positive rate. Often, researchers look at the area under the ROC curve (AUC) to test the predictive capabilities of a model, or compare across multiple models. "The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance." (Fawcett, p 868). For instance, the y=x curve has an area of .5, which again means making a right prediction half the time. A perfect prediction model has an area of 1. In general, the higher the AUC the more accurate the model is in classifying.

A third test of fit suggested by Desjardins is the Brier Score. The Brier score is a unitless index of predictive accuracy (Desjardins, 2002), where $p_i$ is the predictive probability of matriculation for a $individual_i$ and $matriculation_i$ is the actual value of matriculation in the validation sample.

$$Brier\ Score = \frac{\sum[(p_i - matriculation_i)^2]}{n}$$

The Brier score ranges from 0 to 1, and the lower the score the better, as it indicates those with high probabilities do matriculate and those with low predicted probabilities do not. Furthermore, the Brier score measures certainty in forecasts. It could be the case that two models are identical in the ranking of predicted probabilities for applicants but differ in scale. For instance, predicted probabilities for one model could range from 0 to 0.8 while the other model could range from 0 to 1. The Brier Score would prefer the latter, as it is more certain or more committing.

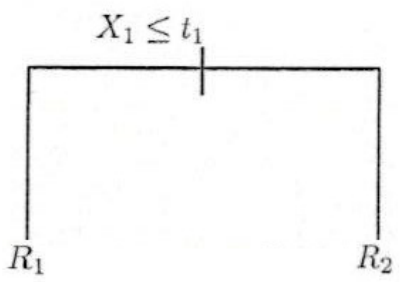
After splitting the samples into training and validation, and using the three aforementioned measures of fit, the parametric model was thoroughly evaluated for its predictive capability.

## 3.4 Non-Parametric Approach: Gradient Boosting

Gradient boosting methods are powerful prediction techniques that have been shown to perform well in many applications. One advantage of using gradient boosting, or non-parametric models in general, is that they are less restrictive. Contrary to parametric models which rely on a researcher specific model design, based on theory or past literature, there is no need to define a model. Because of the lack of restraints imposed, results from machine learning analysis can provide insights into potential unthought-of relationships among variables.

This study appears to be the first of its kind in using gradient boosting to predict college matriculation decisions. The modeling approach is *supervised*, in which the response variable is selected: whether an applicant matriculates or not. The first step in gradient boosting is to select a classifier that will be used to predict an outcome. For instance, an example of a classifier could be the flip of a coin to decide if an applicant will matriculate or not. Rather than randomly guessing outcomes, however, gradient boosting typically uses decision trees.

**Figure 7: Example of a decision tree with depth 1**

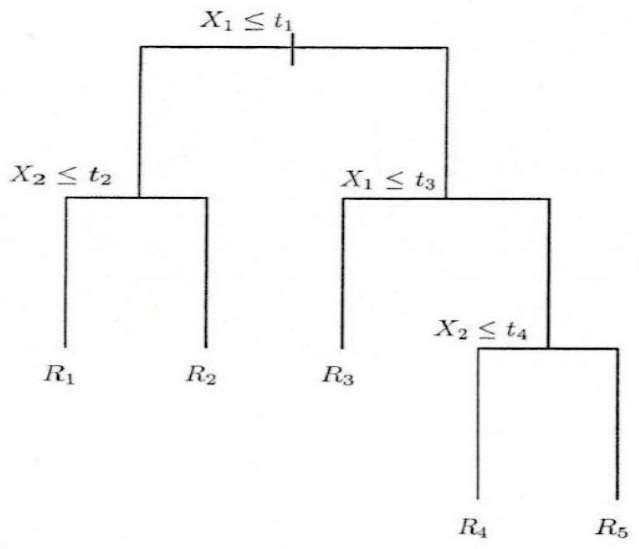



Decision trees essentially look for variables—known as predictors— that split the data into clusters, where clear distinctions can be made among categories of the response variable. For decision trees, the response variable must be binary or categorical, but the predictors can be continuous or count variables. For instance, in figure 7, the sample is split from variable $X_1$,

creating clusters $R_1$ and $R_2$. Provided $X_1$, values less than $t_1$ are classified in $R_1$ while values greater than $t_1$ are classified in $R_2$. Decision trees that have one split , and thus two terminal nodes, are called *stumps*: "A special case of a decision tree with only one split is called a tree stump. In many practical applications, small trees and tree stumps provide considerably accurate results" (Natekin & Knoll, p.7).

Alternatively, one can increase the *depth* of a tree, or the number of splits in each decision tree. For instance, figure 8 below shows a decision tree with depth 3. As the depth of a tree increases, multiple splits from several variables can occur. Having a model with higher depth can shed light into possible interactions among variables. In the context of this study, higher depth trees could indicate which variables grouped together make a good distinction between applicants who matriculate and applicants who do not. It is important to note that a decision tree might split over the same variable more than once, as indicated in the figure below.

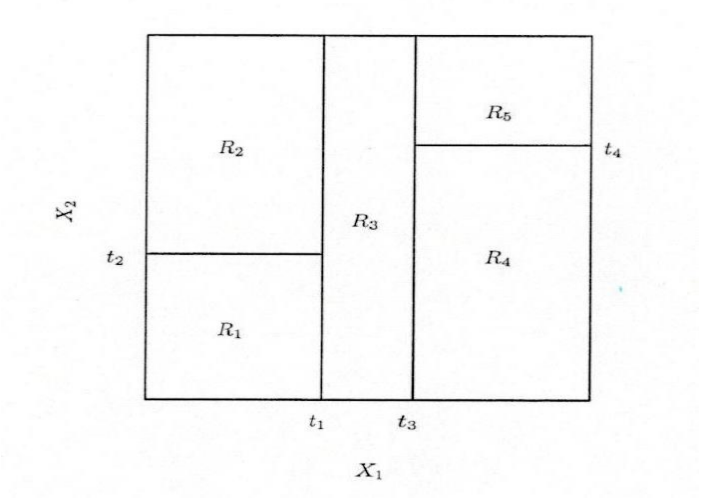
**Figure 8: Example of a decision tree (depth 3)**




Source: Gareth, James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 2014.

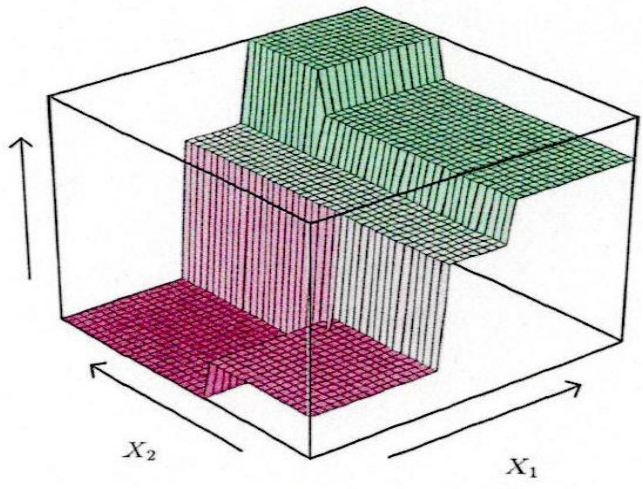
Borrowing from *Introduction to Statistical Learning*, Figure 9 shows clusters $R_1$-$R_5$ created from the decision tree. Correspondingly, figure 10 shows a three-dimensional perspective plot of the regions created by the splits.

**Figure 9: Output from decision tree**




Source: Gareth, James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 2014.

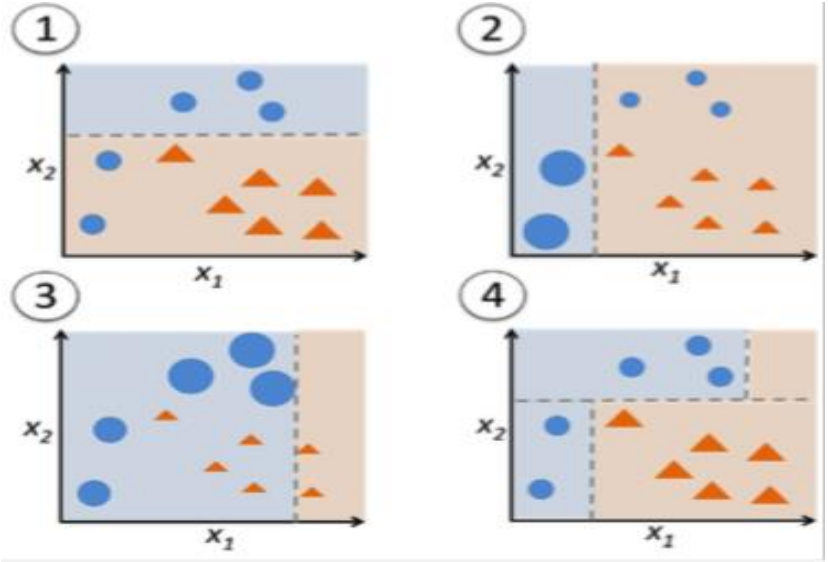
**Figure 10: Perspective plot of prediction surface from decision tree**




Source: Gareth, James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 2014.

Gradient boosting comes from the notion of additive models, where the algorithm will use a classifier—a decision tree in this case— to predict outcomes and sequentially add models correcting previous errors. "The main idea of boosting is to add new models to the ensemble of models sequentially. At each particular iteration, a new weak, base-learner model is trained with respect to the error of the whole ensemble learnt so far" (Natekin&Knoll, p.1).

**Figure 11: Illustration of GBM algorithm**




Source: Sebastian Raschka. "How does the random forest model work? How is it different from bagging and boosting in ensemble models? Nov 1, 2015. URL: https://www.quora.com/How-does-the-random-forest-model-work-How-is-it-different-from-bagging-and-boosting-in-ensemble-models

Figure 11 above is based on classifying a response variable that is binary—either a blue circle or orange triangle— using a tree of depth 1. In the illustration above, a decision tree begins by making a horizontal split of the sample based on a particular value of $x_2$. Orange regions correspond to triangles, and after the initial split two items are incorrectly classified. The two blue circles that got misclassified in the first iteration (known as errors or residuals) are given a

higher weight in stage two, where the algorithm tries to correct itself and adds a second model putting more emphasis in classifying the two misclassified circles. In the next iteration, a split is made for a particular value of $x_1$, where three circles are now misclassified and again receive a higher weight in the next iteration. The gradient boosting algorithm will work to correct these errors by adding yet another model. "In gradient boosting machines, or simply, GBM's, the learning procedure consecutively fits new models to provide a more accurate estimate of the response variable" (Natekin&Knoll, p.1). The additive process correcting residuals continues until step 4, where each item has been classified correctly. The blue regions correspond to circles and orange region correspond to triangles.

While the illustration above is a simplified two-dimensional version of the algorithm, it represents the basic idea behind gradient boosting. In the context of this study, gradient boosting will be used to try to correctly classify applicants into matriculates and non-matriculates, using demographic, ability, timing, and recruiting variables. It is important to note that when using the gradient boosting algorithm, it is possible to accurately classify every single observation. However, the resulting model would essentially be ineffective for predicting out of sample, as it would be too specific to the sample from which it was derived. Therefore, it is important to choose the right place to stop the algorithm to avoid overfitting.

To identify the correct place to stop the algorithm a k-fold cross-validation strategy was used. Also known as a "leave-one-out" approach, the training sample is randomly divided into k folds, 1 to k, where decision trees are used to estimate a model for all but the kth fold (An Introduction to Statistical Learning with Applications in R, p.309 ). The model is validated on the kth fold and the classification error is calculated, with the process being repeated with replacement many times. The algorithm will stop once the error on the held-out sample—also

known as out-of-bag error— is minimized. If the improvement in classification capability from one iteration to the next is small, the algorithm will stop iterating. This concept is also known as shrinkage, or the learning rate of the model: “Similar to a learning rate in stochastic optimization, shrinkage reduces the influence of each individual tree and leaves space for future trees to improve the model” (Chen & Guestrin).

The gbm package in R, version 3.3.1, was used to perform the machine learning analysis. The same training sample utilized for the logistic regression was used as a training sample in gradient boosting. For this study, a five-fold cross validation was used, meaning one fifth of the training sample was randomly selected and held out for internal validation. Models with a combination of different depths values were examined. Classification tables as well as ROC curves were calculated to evaluate the performance of gradient boosting relevant to how well it classified applicants.

Lastly, the non-parametric gradient boosting also provides a measure of variable importance. Variable importance tables indicate which variables in the model played a role in splitting and what percentage of the time each variable was used when classifying applicants. Given that a parametric logistic model is also used in this study, both approaches can be jointly examined to add robustness to results.

# `Chapter 4: Results

The primary question to be addressed in this study is what factors influence matriculation decisions of incoming freshman in the College of Agriculture and Life Sciences (CALS). Matriculation decisions are modeled both parametrically and non-parametrically. The results from both modeling approaches are presented in this chapter.

## 4.1. Parametric Results

### 4.1.1 Descriptive Statistics

Table 3 on page 53 provides a list of the variables used in the parametric logistic regressions. As mentioned in the previous chapter, campus visits and Arizona Experience (AZX) were only tracked from 2013-2015. Thus, a separate analysis is conducted on these later three cohorts to examine the impact of recruiting on matriculation decisions. As far as demographics, an overwhelming majority (82%) of applicants come from the West region of the United States. Approximately sixty percent of applicants are White Americans, and a little over a third are the first in their family to go to college. A very small number of applicants are award recipients and only about five percent of all students applied to the honors college. Relevant to ability, over eighty percent of all applicants submitted a standardized test score, with a mean ACT (or corresponding converted SAT) of 22.5. Moreover, applicants have an average of about two advanced placement units from high school.

As mentioned before, about three quarters of CALS applicants are female students. The most popular majors in CALS are Veterinary and Nutritional Sciences, followed by Animal Sciences. Close to eighty percent of all applicants appear to come from public schools, with the

typical high school size being around 2,000 students. It appears applicants are not the only ones from their high school applying to UA, and have two peers on average applying to a CALS major. As far as recruiting, only ten percent of applicants appear to have attended a campus visit and about four percent attended an AZX visit.

*Truncation in financial aid data*

After careful examination, it appears financial aid information in the data used in this study was truncated, and financial aid offers—both institutional and federal—were observed almost exclusively for students who matriculated. Only 26 of 3,349 applicants who did not matriculate had financial aid data. In contrast, 80% of applicants who matriculated had financial aid data. Incomplete financial aid data would bias the analysis because using a truncated sample would make it seem as if an applicant received financial aid they would almost certainly matriculate. Moreover, a model conducted using this data would most likely overestimate the impact of financial aid, while possibly undermining the effect of other variables in the model. While financial aid is crucial for a comprehensive study of matriculation, one needs complete aid information for both those who matriculated and those who did not. The university ought to keep all records of financial aid, even if a student does not matriculate, to be able to analyze the impact of financial aid on matriculation decisions.

**Table 3: Descriptive Statistics**

| | Variable | N | Mean | Median | 75th | 90th | Minimum | Maximum | Sum |
|---|---|---|---|---|---|---|---|---|---|
| | Matriculation | 5,186 | 0.35 | 0 | 1 | 1 | 0 | 1 | 1,837 |
| *Demographics* | | | | | | | | | |
| | Northeast | 5,186 | 0.05 | 0 | 0 | 0 | 0 | 1 | 275 |
| | Midwest | 5,186 | 0.06 | 0 | 0 | 0 | 0 | 1 | 300 |
| | South | 5,186 | 0.06 | 0 | 0 | 0 | 0 | 1 | 326 |
| | West (Reference) | 5,186 | 0.81 | 1 | 1 | 1 | 0 | 1 | 4,220 |
| | White (Reference) | 5,186 | 0.60 | 1 | 1 | 1 | 0 | 1 | 3,132 |
| | Black | 5,186 | 0.04 | 0 | 0 | 0 | 0 | 1 | 210 |
| | Hispanic_Mexican | 5,186 | 0.07 | 0 | 0 | 0 | 0 | 1 | 365 |
| | Asian | 5,186 | 0.02 | 0 | 0 | 0 | 0 | 1 | 115 |
| | Other_Ethn | 5,186 | 0.18 | 0 | 0 | 1 | 0 | 1 | 912 |
| | Male (Reference) | 5,186 | 0.23 | 0 | 0 | 1 | 0 | 1 | 1,199 |
| | Female | 5,186 | 0.77 | 1 | 1 | 1 | 0 | 1 | 3,987 |
| | d_First_Generation | 5,186 | 0.37 | 0 | 1 | 1 | 0 | 1 | 1,907 |
| | Median_Household_1000 | 5,070 | 72.03 | 70.85 | 88.83 | 113.14 | 0.45 | 216.90 | 365,186.77 |
| *Ability* | | | | | | | | | |
| | d_High_Ability_Student | 5,186 | 0.02 | 0 | 0 | 0 | 0 | 1 | 100 |
| | d_Honors_Admit | 5,186 | 0.05 | 0 | 0 | 0 | 0 | 1 | 269 |
| | ACT_Max | 4,247 | 22.5 | 22 | 25 | 28 | 7 | 36 | 95,552 |
| | AP_Units | 5,186 | 1.94 | 1 | 3 | 5 | 0 | 21.5 | 10,082.09 |
| *Major* | | | | | | | | | |
| | d_ABEMB | 5,186 | 0.03 | 0 | 0 | 0 | 0 | 1 | 150 |
| | d_AGTE | 5,186 | 0.02 | 0 | 0 | 0 | 0 | 1 | 95 |
| | d_ASC | 5,186 | 0.14 | 0 | 0 | 1 | 0 | 1 | 749 |
| | d_ENV | 5,186 | 0.11 | 0 | 0 | 1 | 0 | 1 | 564 |
| | d_MICR | 5,186 | 0.06 | 0 | 0 | 0 | 0 | 1 | 300 |
| | d_NTR | 5,186 | 0.02 | 0 | 0 | 0 | 0 | 1 | 88 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | d_NUSC | 5,186 | 0.23 | 0 | 0 | 1 | 0 | 1 | 1,203 |
| | d_PLS | 5,186 | 0.01 | 0 | 0 | 0 | 0 | 1 | 40 |
| | d_PRFS | 5,186 | 0.05 | 0 | 0 | 0 | 0 | 1 | 248 |
| | d_PRRC | 5,186 | 0.06 | 0 | 0 | 0 | 0 | 1 | 321 |
| | d_VSC (Reference) | 5,186 | 0.26 | 0 | 1 | 1 | 0 | 1 | 1,370 |
| | d_Other_Major | 5,186 | 0.01 | 0 | 0 | 0 | 0 | 1 | 58 |
| *Year Dummies* | | | | | | | | | |
| | d_2011 | 5,186 | 0.17 | 0 | 0 | 1 | 0 | 1 | 903 |
| | d_2012 | 5,186 | 0.21 | 0 | 0 | 1 | 0 | 1 | 1,070 |
| | d_2013 | 5,186 | 0.21 | 0 | 0 | 1 | 0 | 1 | 1,114 |
| | d_2014 | 5,186 | 0.19 | 0 | 0 | 1 | 0 | 1 | 1,000 |
| | d_2015 (Reference) | 5,186 | 0.21 | 0 | 0 | 1 | 0 | 1 | 1,099 |
| *Timing of Application* | | | | | | | | | |
| | Days_ahead_app | 5,186 | 270.35 | 281 | 315 | 343 | 17 | 392 | 1,402,034 |
| *High School Characteristics* | | | | | | | | | |
| | HS_Size | 4,710 | 1,628.64 | 1,689 | 2,172 | 2,740 | 10 | 4,830 | 7670900 |
| | d_Magnet | 4,723 | 0.03 | 0 | 0 | 0 | 0 | 1 | 119 |
| | d_Catholic | 4,723 | 0.1 | 0 | 0 | 0 | 0 | 1 | 460 |
| | d_Charter | 4,723 | 0.04 | 0 | 0 | 0 | 0 | 1 | 172 |
| | d_Public_School(Reference) | 4,723 | 0.86 | 1 | 1 | 1 | 0 | 1 | 4,062 |
| | d_Private_School | 4,723 | 0.14 | 0 | 0 | 1 | 0 | 1 | 661 |
| | HS_Peers | 5,085 | 2.11 | 1 | 3 | 6 | 0 | 16 | 10,724 |
| | *Past_Peers | 5,085 | 4.12 | 1 | 5 | 13 | 0 | 46 | 20,935 |
| ***Recruiting* | | | | | | | | | |
| | Campus_Tour | 5,186 | 0.08 | 0 | 0 | 0 | 0 | 1 | 394 |
| | AZX_Visit | 5,186 | 0.03 | 0 | 0 | 0 | 0 | 1 | 159 |

*Measured for 2012-2015 **Measured 2013-2015

### 4.1.2 Logistic Model Results

The overall sample was randomly divided into a training set and a validation set. The training set was used to estimate parameter coefficients while the validation tested how well the training model classified applicants. The training set consisted of seventy percent of the sample while the validation sample consisted of the remaining thirty percent. The procedure *Surveyselect* in SAS was used to randomly partition the data, and the sample was stratified by academic cohort to ensure each year's applicants were represented proportionally. Stratifying the sample was important to make sure the training set was representative of the overall sample, and had a proportionate count of each cohort from 2011-2015. Although the procedure used to divide the sample is random, a test of difference in means to check for balance between training and validation samples can be found in appendix 4. It is important to note that when splitting the dataset using the procedure *Surveyselect*, one would obtain different samples depending on the specified seed value. The reason model results might be sensitive to a seed value is that some variables in the model—such as AZX visits or high ability students for instance— have a small number of occurrences. The variables with low occurrences pose a potential issue, as the split is susceptible to randomly, yet disproportionate number of students with these characteristics in training and validation. To retain the integrity of this procedure, it is important to pick a seed that guarantees a sample with equal proportions of each variable in training and validation, particularly for those variables with low occurrences.

The marginal effects presented here were computed using Stata 14, and were obtained from the training sample alone. Marginal effects were calculated at sample means. For dummy variables, marginal effects were calculated as discrete changes in probability. Furthermore, as noted in Table 4, the reference groups for dummy variables were applicants from the West,

White , 2015 cohort , veterinary science majors, and applicants from public schools. Table 4 on page 60 provides the estimated marginal effects of three different model specifications. Model 1 includes cohorts 2011-2015 with all aforementioned variables excluding past peers. Since the data in this sample only goes back to 2011, there is no measure of previous peers for this cohort. Model 2 includes cohorts 2012-2015 to examine the effect of past peers. Model 3 includes cohorts 2013-2015 to examine the influence of recruiting variables on matriculation.

In all three models, it appears that applicants from the South and Midwest are about 15% and 9%, respectively, less likely to matriculate than applicants from the West. This finding could be due to more competition in agricultural program availability at nearby land-grant universities in both the South and Midwest. Median household income was negative and significant in all three model specifications as well, though the magnitude is quite small. An additional ten thousand dollars in median household income decreases the probability of matriculation by around one percent. Applicants who were admitted to the Honors College were 70% more likely to matriculate. While this effect could be picking up the fact that honors admits typically receive more financial aid, perhaps the smaller classes, private dorms, and research opportunities provided by the honors college are enticing perks for incoming students.

Relevant to ability, there appears to be a concave influence of standardized test scores on matriculating decisions. In all three models, ACT scores are positive, but the square of ACT scores is negative and significant. This result indicates that as a student's score increases, they are more likely to matriculate, though at a decreasing rate. This result is intuitive, as UA is not a particularly selective institution and students with high standardized test scores are more likely to have numerous alternative offers from other schools. The sign on the number of AP units is negative, though not statistically significant. As far as majors, compared with veterinary

sciences, environmental science majors are about 9% less likely to matriculate, while retailing and family-studies applicants are more likely to matriculate, 9% and 15% respectively.

An interesting result of this analysis that the timing of an application appears to be a robust indicator of matriculation intentions. The days-ahead variable remained significant in all three model specifications, indicating the earlier a student applies to the university the less likely he or she is to matriculate. Every week away from the first day of classes decreases the probability of matriculation by around 2%. Early applications do not correspond to an early intention or preferred institution as originally thought. Rather, it appears if students apply to UA early, then they are probably applying to other schools early as well. Furthermore, the squared term of days ahead is positive—though the magnitude is small—, meaning the probability of matriculation decreases at an increasing rate. Figure 12 displays the predicted probability of matriculation for both an average-scoring and high-scoring applicant (90$^{th}$ percentile).

**Figure 12: Predicted probabilities for two types of applicants**

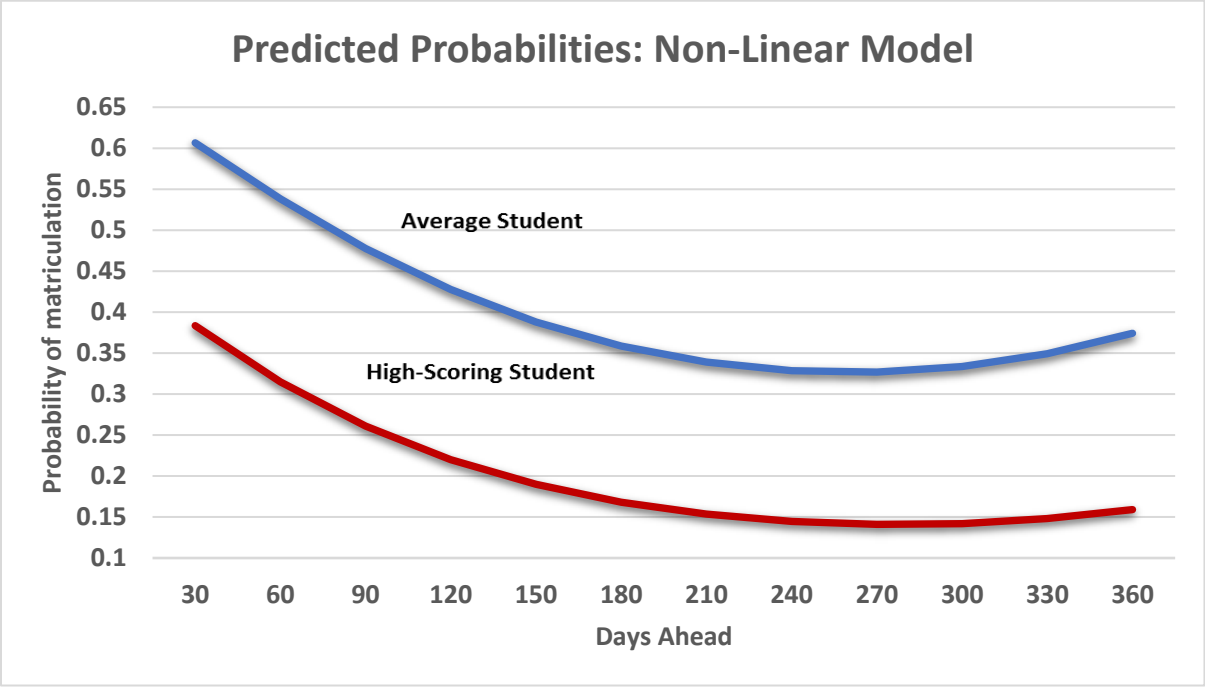
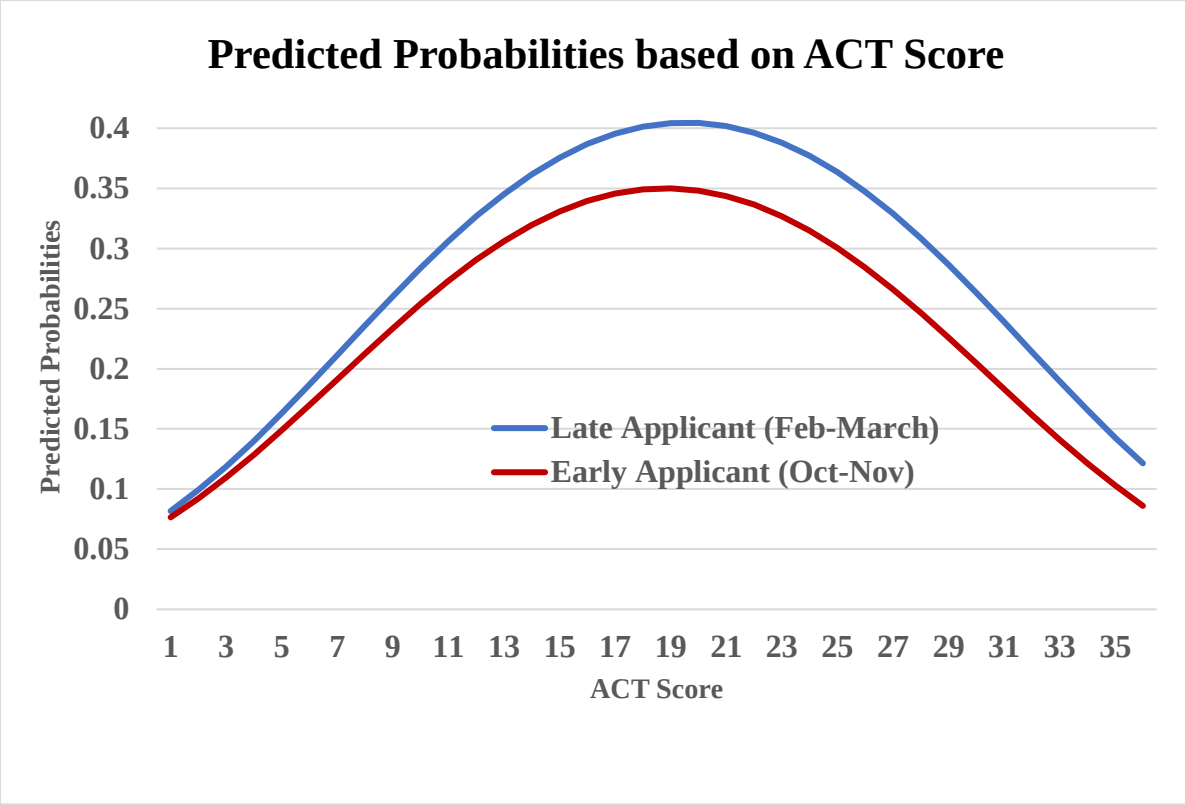


Figure 12 indicates that as the days ahead increases, or the earlier a student applies, the probability of matriculation decreases. Moreover, this relationship holds for both average students and high-scoring students, but high scoring students are significantly less likely to matriculate at any days-ahead value. Similarly, figure 13 shows the predicted probabilities of a typical applicant in CALS based on different ACT scores. As the ACT score of an applicant increases, so does his probability of matriculation, reaching a maximum between 19 and 20. However, once applicants reach a score higher than 20, the probability of matriculation decreases the higher an applicant scores[3].

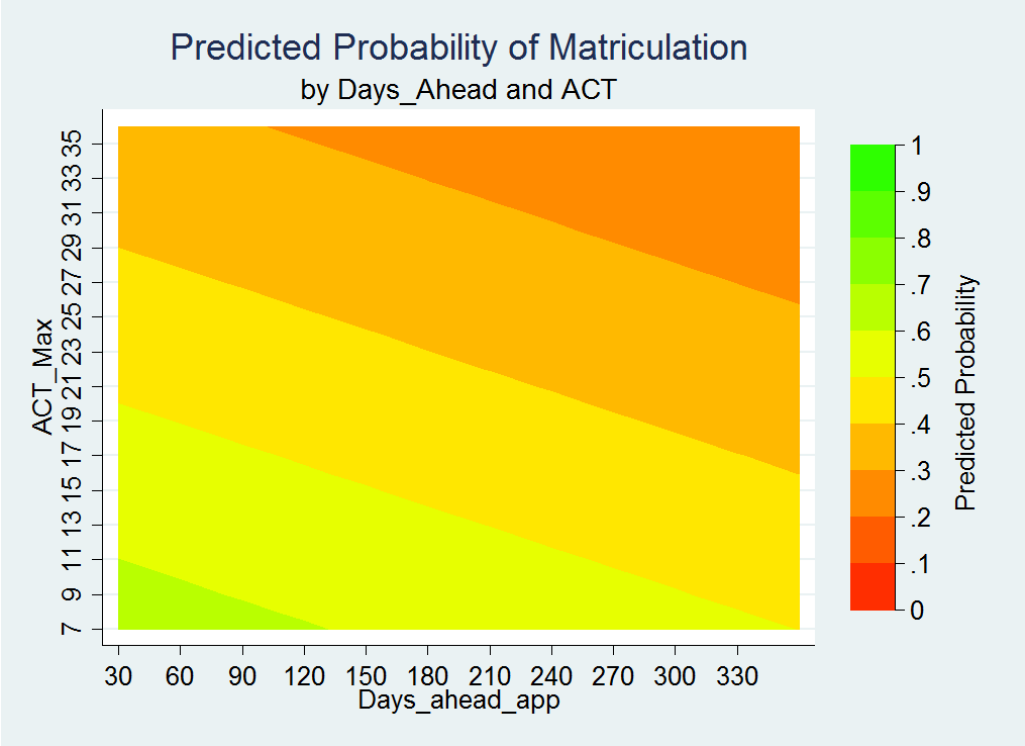
**Figure 13: Predicted probabilities based on ACT score**




[3] Applicants who did not submit standardized test scores were excluded from the analysis. However, models including students with no ACT or SAT scores—using a dummy variable for having a test score times the actual score—can be found in appendix 6. Results are consistent with or without these applicants.

Figure 14 below shows a two-way contour plot displaying predicted probabilities from a model with only linear terms of days-ahead and ACT scores included in the model. To generate the contour plot, regression coefficients were multiplied by sample means for continuous variables and by the mode for dummy variables. Predicted probabilities were calculated at different combinations of ACT scores and days ahead. Thus, this contour plot shows matriculation propensities for a representative candidate in CALS.

**Figure 14: Contour Plot of Predicted Probabilities based on a Linear Model**




For every level of ACT, the probability of matriculation decreases the earlier the student applies. Applicants who score low and apply last minute are the most likely to matriculate. Conversely, high scoring students that apply early to the university are the least likely to matriculate.

None of the high school-specific demographics turned out to be significant predictors of matriculation. However, both peer effects are statistically significant, though in model 3 only the past peers variable was significant. For models 1 and 2, every additional peer in the same graduating class applying to a CALS major raises the probability of matriculation between one and two percent. When the past peer variable was added in model 2, it was positive and significant as well, though the magnitude of this effect was about a third of the effect of current peers. Both results shed light into potential "flock" and "norm" influences. In other words, students appear to apply in groups, and the more peers applying to the UA —and similar program since these are all CALS majors—a student has, the more likely they themselves are to matriculate. This finding could be attributed to students grouping together or feeling more comfortable matriculating at a place with friends or acquaintances. As far as "reputation" effects, it also appears that students are influenced by application and matriculation decisions of previous students.

Relevant to recruiting, both campus visits and AZX are significant and positive. Those who attended a campus tour were about 26% more likely to matriculate while those who attended an AZX were 22% more likely to matriculate. While the difference in magnitude may or may not be statistically significant, this result is interesting, as AZX visits are usually longer events and cost more to run. The fact that recruiting appears to increase the probability of matriculation is good news for recruitment managers. However, a further issue to investigate would be self-selection. Are students who attend campus visits more likely to matriculate because recruiting 'wins them over'? Or is because the students that attend campus visits already have a high propensity to matriculate at UA in the first place?

**Table 4: Marginal effects**

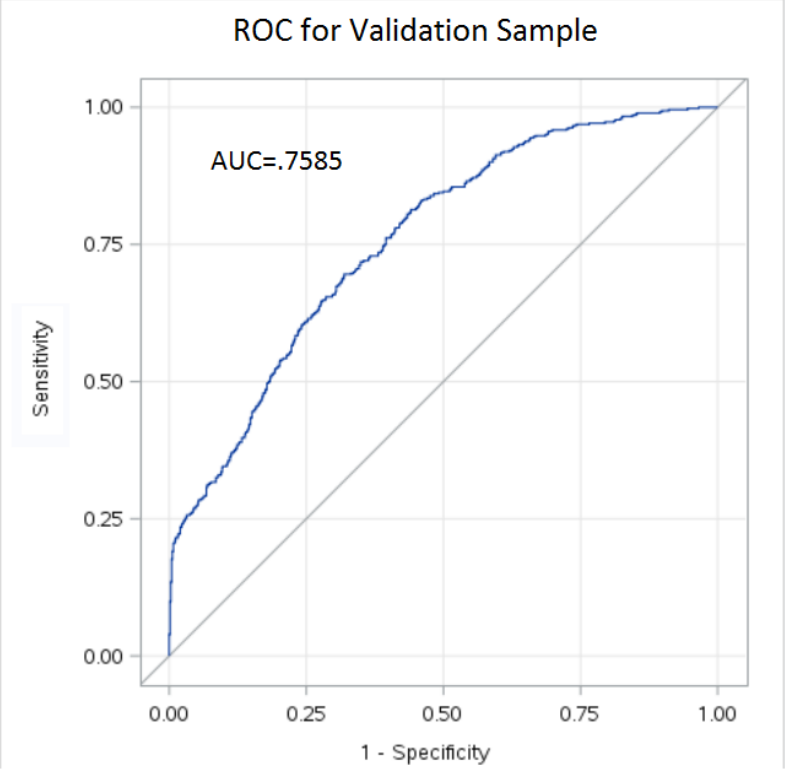
| | Model 1 (N=2,713) No past peers & no recruitment | | Model 2 (N=2,302) No recruitment | | Model 3 (N=1,779) | |
|---|---|---|---|---|---|---|
| | **Marginal Effect** | **Z** | **Marginal Effect** | **Z** | **Marginal Effect** | **Z** |
| ***DEMOGRAPHICS*** | | | | | | |
| Northeast | -0.0447 | -1.07 | -0.0530 | -1.18 | -0.0519 | -1.04 |
| Midwest | -0.0979 | -2.2 | -0.1091 | -2.22 | -0.1067 | -1.93 |
| South | -0.1521 | -3.1 | -0.1693 | -3.19 | -0.1644 | -2.8 |
| Black | -0.0754 | -1.49 | -0.0544 | -1 | -0.0425 | -0.71 |
| Hispanic_Mexican | 0.0059 | 0.17 | 0.0266 | 0.71 | 0.0516 | 1.31 |
| Asian | 0.0461 | 0.75 | 0.0334 | 0.5 | 0.0119 | 0.16 |
| Other_Ethn | -0.0004 | -0.02 | 0.0018 | 0.07 | 0.0185 | 0.66 |
| Female | -0.0188 | -0.81 | -0.0273 | -1.08 | -0.0597 | -2.19 |
| d_First_Generation | -0.0138 | -0.7 | -0.0034 | -0.16 | 0.0046 | 0.2 |
| Median_Household_1000 | -0.0012 | -3.81 | -0.0011 | -3.12 | -0.0012 | -3.12 |
| ***ABILITY*** | | | | | | |
| d_High_Ability_Student | -0.0978 | -1.1 | -0.1137 | -1.17 | -0.2133 | -1.72 |
| d_Honors_Admit | 0.7210 | 11.48 | 0.7325 | 10.15 | 0.7276 | 8.72 |
| ACT_Max | 0.0407 | 2.16 | 0.0399 | 1.88 | 0.0484 | 2.08 |
| ACT_SQ | -0.0011 | -2.86 | -0.0012 | -2.71 | -0.0013 | -2.73 |
| AP_Units | -0.0010 | -0.24 | 0.0004 | 0.08 | 0.0021 | 0.45 |
| ***MAJOR*** | | | | | | |
| d_ABEM | -0.0126 | -0.23 | -0.0242 | -0.42 | -0.0658 | -1.09 |
| d_AGTE | 0.0143 | 0.22 | -0.0251 | -0.36 | 0.0181 | 0.25 |
| d_ASC | -0.0436 | -1.5 | -0.0654 | -2.07 | -0.0347 | -1 |
| d_ENV | -0.0969 | -2.7 | -0.1409 | -3.52 | -0.1190 | -2.7 |
| d_MICR | 0.0295 | 0.74 | 0.0279 | 0.65 | 0.0227 | 0.48 |
| d_NTR | 0.0387 | 0.56 | 0.0070 | 0.09 | -0.0077 | -0.09 |

| | Marginal Effect | Z | Marginal Effect | Z | Marginal Effect | Z |
|---|---|---|---|---|---|---|
| d_NUSC | 0.0133 | 0.54 | -0.0074 | -0.28 | -0.0138 | -0.48 |
| d_PLS | 0.0789 | 0.69 | 0.0071 | 0.05 | -0.2270 | -1.19 |
| d_PRFS | 0.0886 | 2.2 | 0.0358 | 0.81 | 0.0653 | 1.36 |
| d_PRRC | 0.1623 | 4.5 | 0.1450 | 3.7 | 0.1441 | 3.41 |
| d_Other_Major | 0.0357 | 0.44 | 0.0368 | 0.38 | 0.0175 | 0.16 |
| ***YEAR DUMMIES*** | | | | | | |
| d_2011 | -0.0088 | -0.3 | . | . | . | . |
| d_2012 | 0.0527 | 1.97 | 0.0920 | 3.12 | . | . |
| d_2013 | -0.0085 | -0.32 | 0.0171 | 0.62 | 0.0277 | 1.04 |
| d_2014 | -0.0332 | -1.26 | -0.0245 | -0.91 | -0.0158 | -0.62 |
| **TIMING OF APPLICATION** | | | | | | |
| Days_ahead_app | -0.0029 | -2.61 | -0.0028 | -2.27 | -0.0026 | -1.99 |
| Days_app_sq | 0.00001 | 3.02 | 0.000005 | 2.39 | 0.000004 | 2.05 |
| Interaction_act_days_app | 0.000004 | 0.12 | 0.00002 | 0.41 | 0.00001 | 0.14 |
| ***HIGH SCHOOL VARIABLES*** | | | | | | |
| HS_Size | -0.000003 | -0.23 | -0.00002 | -1.38 | -0.00002 | -1.16 |
| d_Magnet | -0.0052 | -0.08 | 0.0153 | 0.24 | 0.0331 | 0.51 |
| d_Catholic | 0.0460 | 0.9 | 0.0406 | 0.73 | 0.0127 | 0.2 |
| d_Charter | 0.0777 | 1.52 | 0.0690 | 1.29 | 0.0865 | 1.51 |
| d_Private_School | -0.0027 | -0.06 | -0.0080 | -0.16 | -0.0171 | -0.3 |
| HS_Peers | 0.0231 | 7.22 | 0.0149 | 3.3 | 0.0052 | 1.01 |
| Past_Peers | . | . | 0.0055 | 3.02 | 0.0079 | 4.19 |
| ***RECRUITING*** | | | | | | |
| Campus_Tour | . | . | . | . | 0.2664 | 9.81 |
| AZX_Visit | . | . | . | . | 0.2302 | 5.43 |

### 4.1.3 Test of fit

The second part of the analysis relates to the predictive capability of the model, particularly for subsequent incoming classes. After parameters were estimated in the training sample, the estimated coefficients were applied to the validation sample. Figure 15 below displays the ROC curve for the validation sample. Since the ROC curve and area under the curve (AUC ) was similar for all three models, only the ROC curve for model 3 is presented. The AUC for this model is 0.75, which signals the model fits the validation sample well. This result is similar to that found by Desjardins who obtained an ROC with AUC of 0.72.

**Figure 15: Receiver operating curve (ROC) for validation sample**




Another way to test the efficacy of the model predictions is with classification tables. Table 5 below describes how well the model correctly classified applicants in the validation

sample. Three different cutoffs were used, and it appears that using 0.35—which is the historical yield rate for CALS in the past five years— produced the most accurate classifications in regards to correctly identifying applicants who matriculate. The model in general produced low predicted probabilities, and it appears that using cutoffs above the historical yield lead to poor predictions regarding the true positive rate.

**Table 5: Classification for validation sample**

| Cutoff | True Positive | False positive | True Negative | False Negative | Overall Accuracy |
|---|---|---|---|---|---|
| 0.35 | 0.70 | 0.30 | 0.68 | 0.27 | 0.690 |
| 0.5 | 0.45 | 0.55 | 0.88 | 0.34 | 0.702 |
| 0.65 | 0.30 | 0.70 | 0.97 | 0.35 | 0.695 |

Lastly, in table 6 the validation sample was divided into deciles, ranging from greatest to lowest by number of matriculates. Here no cutoff values are used. Rather, for each decile, the average predicted probability was used to estimate how many applicants would matriculate in that decile. Predictions are obtained by multiplying the average predicted probability times the number of applicants within each decile, producing an estimated yield. For example, in group 5, there were 28 students who matriculated and the model predicted 25, so a difference of three students. It is interesting that the model appears to predict better at the lower deciles, where there are fewer applicants who matriculate. Perhaps more accurate prediction for lower deciles could be attributed to the fact that the sample consists of mostly students who did not matriculate, so there are more observations to discern patterns of non-matriculation. Finally, the Brier score is 0.19, indicating that the model produces accurate forecasts. This is also similar to Desjardins, who obtained a Brier score of 0.21.

**Table 6: Prediction in validation sample (by decile)**

| | | Matriculated | | Did not Matriculate | |
|---|---|---|---|---|---|
| Group | Total | Actual | Predicted | Actual | Predicted |
| 1 | 71 | 64 | 62.3 | 7 | 8.7 |
| 2 | 71 | 43 | 44.5 | 28 | 26.5 |
| 3 | 72 | 36 | 35.3 | 36 | 36.7 |
| 4 | 71 | 30 | 28.6 | 41 | 42.4 |
| 5 | 72 | 28 | 25.0 | 44 | 47.0 |
| 6 | 71 | 33 | 21.2 | 38 | 49.8 |
| 7 | 71 | 27 | 18.1 | 44 | 52.9 |
| 8 | 72 | 14 | 15.1 | 58 | 56.9 |
| 9 | 71 | 12 | 11.6 | 59 | 59.4 |
| 10 | 72 | 5 | 6.2 | 67 | 65.8 |
| | Brier Score | 0.1921 | | | |

## 4.2 Non-Parametric Results

In order to make a valid comparison between the parametric and non-parametric analysis, the same training and test samples were used in both logistic models and gradient boosting. Gradient boosting was used to obtain predicted probabilities of matriculation, making it possible to compare the different approaches in how well they classified applicants. Gradient boosting results are presented for the cohorts of 2013-2015 to examine influence of recruitment and past peers.

As mentioned, the only tuning parameter that varied across gradient boosting models was interaction depth, which is the number of splits in the decision tree allowed at each iteration. Varying the interaction depth can possibly reveal insights into interactions of variables and nonlinearities that together indicate whether an applicant matriculates. Four different depths were

examined to study sensitivity and robustness of results. Shrinkage value was held constant at .005, and a five-fold cross validation approach was used.

### 4.2.1 Variable Importance

A useful feature of gradient boosting is that it provides a measure of variable importance. Variable importance in a non-parametric model is analogous to a variable being significant in a parametric model. In gradient boosting, variables are ranked by how often—or the percentage of all iterations—that each variable was used in creating a split. In other words, which variables were used to make a distinction between applicants who matriculate and applicants who do not. Table 7 below shows the ranking of the top eight variables from most important to least important identified by the different gradient boosting models across four different interaction depths. Additionally, Table 7 shows the number of iterations that minimized cross-validation error for each model.

**Table 7: Ranking of top variables used for splitting**

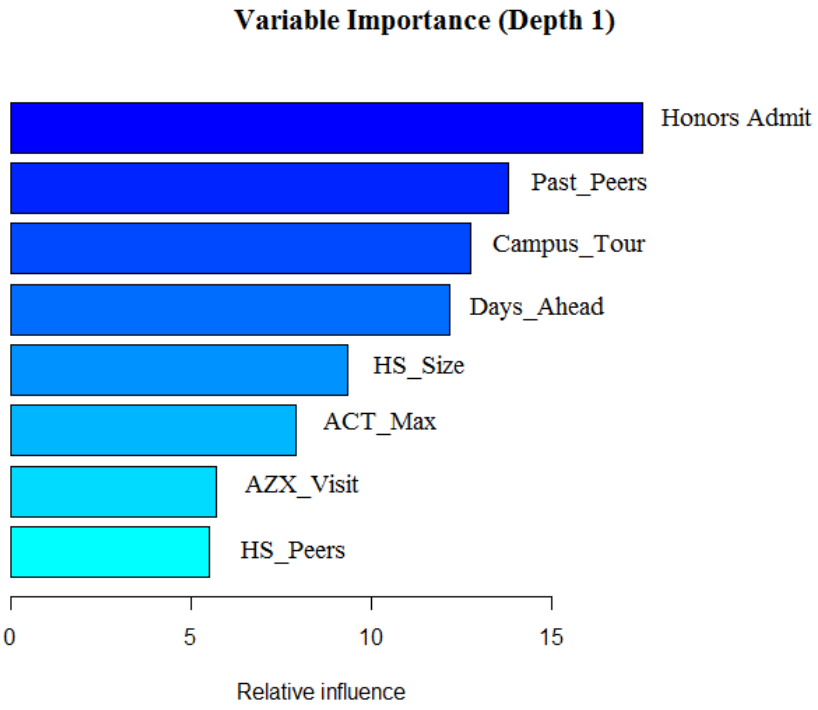
| | Depth 1 | Depth 2 | Depth 3 | Depth 4 |
|---|---|---|---|---|
| | d_Honors_Admit | Days_ahead_app | Days_ahead_app | Days_ahead_app |
| | Past_Peers | HS_Size | HS_Size | HS_Size |
| | Campus_Tour | Past_Peers | Past_Peers | Past_Peers |
| | Days_ahead_app | d_Honors_Admit | ACT_Max | ACT_Max |
| | HS_Size | ACT_Max | d_Honors_Admit | d_Honors_Admit |
| | ACT_Max | Campus_Tour | Campus_Tour | HS_Peers |
| | AZX_Visit | HS_Peers | HS_Peers | Campus_Tour |
| | HS_Peers | AZX_Visit | AP_Units | AP_Units |
| Iterations | 2,929 | 1,578 | 1,047 | 1,079 |

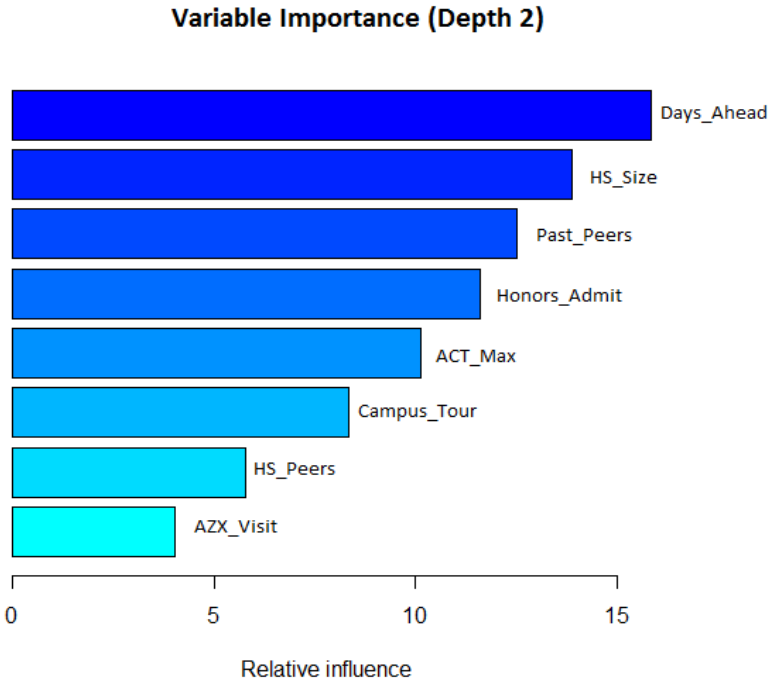
From a model with an interaction depth of one, honors applicants, past peers, campus tours and days ahead were used most frequently in creating splits. These variables were also significant in the logistic regressions and had large marginal effects. As shown on the table,

there is not much fluctuation with regards to the ranking of variables from depths 2-4. Interestingly, once interaction depths higher than one were introduced, the model identified high school size as an important variable to split on. This result contrasts with the logistic regressions, where high school size was not significant in any model specification.

In addition to variable importance, gradient boosting also provides the relative influence of each variable. Figure 16 presents the relative influence of the top eight variables for depths one and two. Since the ranking and relative influence was similar for depths 2-4, only the relative influence for depth 2 is presented. In depth 1, the model identified honors applicants as the most important variable, and honors_admit was used to create a split around twenty percent of the time. Past peers, campus_tour, and days_ahead were similar in relative influence and were used to create splits thirteen percent in all iterations.

**Figure 16: Relative influence from gradient boosting**



In contrast, for models with depths higher than one, the timing of an application was used to create splits most frequently at around seventeen percent of the time, followed by high school size at around fourteen percent.

### 4.2.2 Accuracy of gradient boosting

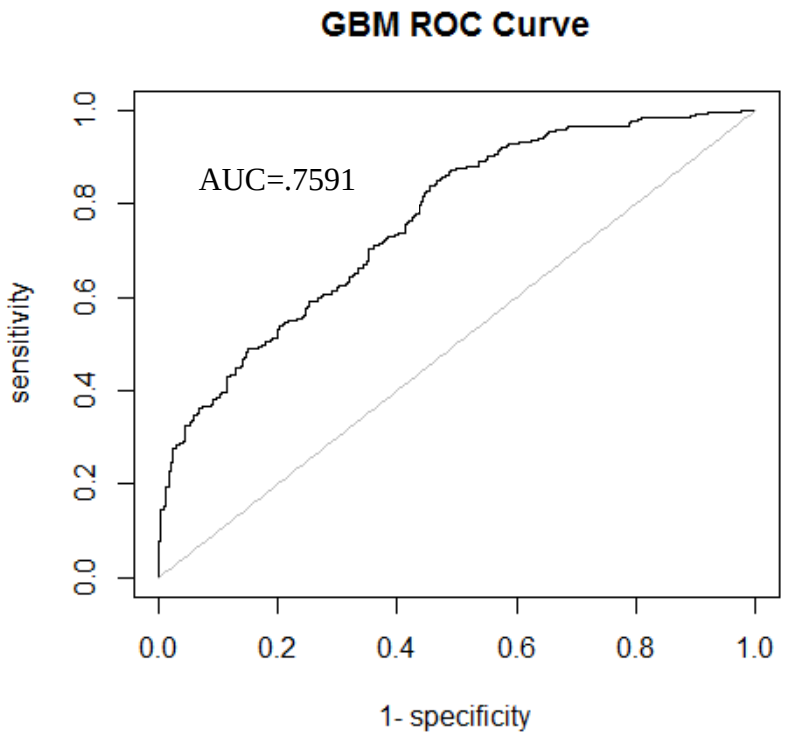
Table 8 below shows a classification table for each model with a designated cutoff of 0.35, the historical yield rate for CALS in the sample period. It appears varying the depth of the model had no significant impact on its performance, as the accuracy percentages remained consistent across all four depths. The true positive rate remained in the low sixties while the true negative rate stayed at seventy. Much like the parametric logistic regressions, gradient boosting appears to be more accurate at predicting applicants who did not matriculate.

**Table 8: Predictive Accuracy of GBM models (cutoff=0.35)**

| Depth | True Positive | False positive | True Negative | False Negative | Overall Accuracy |
|---|---|---|---|---|---|
| 1 | 0.61 | 0.39 | 0.70 | 0.30 | 0.668 |
| 2 | 0.63 | 0.38 | 0.70 | 0.30 | 0.665 |
| 3 | 0.62 | 0.37 | 0.69 | 0.31 | 0.662 |
| 4 | 0.62 | 0.38 | 0.69 | 0.31 | 0.661 |

Similarly, figure 17 below shows the ROC curve from the test sample with an interaction depth of 2. The AUC was .7591, so the model appears to fit the test sample well. Additionally, the AUC is remarkably close to the one obtained from the logistic regressions.

**Figure 17: ROC curve for gradient boosting model (depth 2) validation sample**




### 4.3 Logistic vs Gradient Boosting

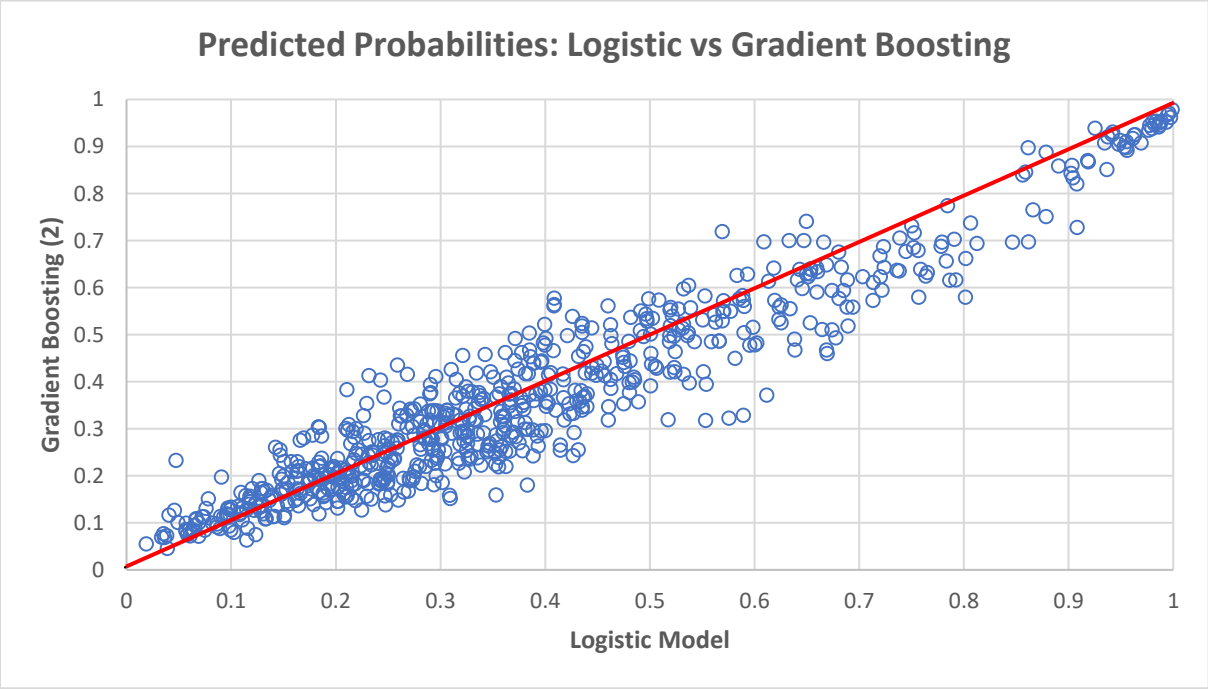
In addition to identifying factors influencing matriculation, a secondary objective of this study is to compare the predictive accuracy of logistic models with machine learning methods. As far as variable importance and significance, both approaches almost coincided in which variables were deemed important to predict matriculation decisions. For the most part, variables that had the largest marginal effects in the logistic regressions had the highest relative influence in gradient boosting. The only exception was high school size, which was identified by gradient boosting as being important, but not the case in the logistic models. Since the size of a high school increased in relative influence for higher depths, a possible hypothesis is that perhaps

high school size interacted with some other variable creates a good distinction between matriculates and non-matriculates.

In general, it appears that the logistic models were more accurate at classifying applicants who did matriculate. The true positive rate of the logit models was 0.70, compared with 0.63 from gradient boosting. Gradient boosting was slightly more accurate in identifying applicants who did not matriculate at a 0.70 true negative rate, compared with 0.68 in the logistic models. The logistic model was slightly more accurate overall, at 0.69 accuracy compared with 0.67 in gradient boosting.

Another way to compare results from the logistic regressions and gradient boosting is by comparing predicted probabilities. Figure 18 below shows a scatterplot of predicted probabilities in the validation sample produced by both approaches.

**Figure 18: Scatter plot of predicted probabilities**



The fact that most points are concentrated near the y=x axis means both models tend to produce similar probabilities. With that said, there appears to be more dispersion in applicants with predicted probabilities on the mid ranges—0.4 to 0.7—, and logistic models tend to produce (slightly) higher predicted probabilities overall. Descriptive statistics for predicted probabilities from both modeling approaches can be found in appendix 4.

# `Chapter 5: Conclusions and Implications

This study sought to investigate factors influencing matriculation decisions for admitted applicants in the College of Agriculture and Life sciences (CALS) at the University of Arizona (UA). Aside from the typical binary response model used in the literature, non-parametric machine learning methods were used to check the robustness of parametric results.

Significant time and resources are spent on attracting new applicants each year while trying to maintain or increase yield rates. Facing potential budget cuts, it is important to allocate resources efficiently and target students groups with the highest payoffs. Studies like these are useful because they can reveal new insights in matriculation trends, leading to beneficial and more effective policy interventions. This study can also serve to corroborate previous intuition of recruiting managers with statistical analysis.

## 5.1 Implications

Findings from this study reveal that the timing of an application is a robust indicator of matriculation decisions. The days-ahead variable remained significant in all logistic model specifications and was identified as the top variable in the gradient boosting analysis. This study found the further ahead a student applies before the first day of classes, the less likely they are to matriculate. Furthermore, students who apply very early are the least likely to matriculate. It appears that early applications do not signal an applicant is fully committed to University, or a "sure bet" as perhaps previously thought. The fact that students who apply early matriculate less often could be that these early applicants might just be more organized, plan ahead, and most likely apply early to many other schools as well. Students who apply at the beginning of the application period also receive offers early and have time to "shop around" for other offers. It

could also be applicants who apply late were rejected by other universities and UA serves as a backup school.

An alternative explanation could be that early applicants lose engagement with the UA along the way. For students that apply in September and October, a lot can happen in the several months between when they apply and the first day of classes. Perhaps these students change their mind about matriculating because they lose that initial excitement about the university. Declining interest over time is something CALS could change or influence at relatively low cost. For example, by simply contacting these early applicants throughout their senior year, enrollment managers can make sure students do not lose interest or forget about all the great opportunities UA has to offer. The goal should be to keep students engaged throughout their admission process and remind them why UA is a great viable option.

An interesting finding of this study is that students' matriculation decisions are influenced by both their high school peers and predecessors. This study found applicants with more peers who were admitted to UA were more likely to matriculate. This finding could be attributed to students tending to group together, or feeling more comfortable matriculating at a place with friends or acquaintances. Moreover, applicants from high schools that have a history of sending students to the UA are more likely to matriculate. Thus, there does appear to be a "past peer" effect, and applicants are indeed influenced by what students in the previous graduating classes have done. Once students apply to a CALS major at the UA, it sets a precedent and opens the door for future students to follow.

It could also be that the peer and past peer variables serve as proxies for unobserved influences. Perhaps high schools where many students apply to the UA have connection to the university in the form of counselors or teachers that keep encouraging students to apply.

Alternatively, perhaps these high schools were exposed to recruiting efforts in the past. Regardless, a good strategy for CALS recruiting appears to be to target high schools more aggressively, as establishing a connection with a high school seems to not only have an immediate positive effect on matriculation but also a long-lasting influence.

Good news for recruiting managers is that campus tours and Arizona Experience (AZX) visits have a positive impact on matriculation. Students who attend either a campus tour or AZX visit are more likely to matriculate. A possible explanation could be self-selection, as students who attend campus visits might already have a high propensity to matriculate. Self-selected students know that UA is their top choice and attend a tour to get to know the university before eventually matriculating. This study did not control for the possible self-selection bias.

Nonetheless, it might very well be that campus visits win students over. The University of Arizona has one of the largest, attractive, diverse campuses in the Southwest. UA has nationally recognized recreational facilities, a 57,000-seat football stadium and top-of-the-line dormitories. The UA also has a strong community, and students who attend a campus visit are educated on school traditions and the millions of dollars spent on research annually. It is not implausible to believe that exposing high school students to such an impressive environment has the potential to make a great impression, drawing students to matriculate.

Recruiting managers in CALS have recently been discussing offering $200-$300 scholarships for students who attend campus tours. Based on findings from this study, this strategy would be a profitable idea and pay dividends, as attending a campus visit greatly increases the probability of matriculation. CALS could also think about flying out specific students for a campus visit, particularly high ability students if the goal is to attract more talent.

As far as using findings from this study for future decision making, both the logistic and

gradient boosting models can be directly applicable to future incoming classes. By mid-January, CALS will have received over 80% of all applicants. This means that upon collecting and cleaning applicant data, most of the variables in this model can be obtained and each applicant could get an estimated probability of matriculation. By January, CALS could use this model to identify applicants "on the fence", or those who have mid-range probabilities of matriculating. From an efficiency perspective, it is not worth going after students with low-range probabilities, as no matter how CALS tries to recruit, these students are unlikely to come. Along the same lines, applicants with high probabilities of matriculation are almost a sure bet, so there is no need to invest additional resources to recruit them. The most efficient way to improve yield rates is to go after students on the fence. Gaining students on the margin can lead to significant increases in yield rates and bring in more talent and revenue within the college.

### 5.2 Data Deficiencies:

Ideally, a study of matriculation should be done at the aggregate university level. While findings from this study can be extremely useful and provide guidance to enrollment managers in CALS recruiting, results may not be applicable to the University as a whole. CALS is the fifth largest college at the University of Arizona, and students who apply to CALS might be systematically different from the rest of the university. Having data on the whole university would not only increase sample size –leading to more robust results— but also make this study applicable to other universities.

A disappointing deficiency of the sample used in this study was that financial aid data was incomplete. As already discussed, financial aid information was available almost exclusively for students who matriculated. Only 26 of 3,349 applicants who did not matriculate had financial aid data, making it very likely that the financial aid data were truncated. Perhaps the university

does not keep an accurate record of financial aid offers for students who do not matriculate.

Financial aid is without a doubt a prominent influencer of matriculation decisions and many studies in the past have analyzed the impact of financial aid. Examining the impact of financial aid on matriculation could lead to direct cost savings in CALS or more efficient campaigns. Furthermore, financial aid is a crucial component that needs to be controlled for as otherwise, the influence of other variables gets confounded. When the truncated financial data was included in the model, the significance and sign of all the key variables remained the same, but the magnitude of them shrank significantly. This means that the marginal effect of some of the key variables in this study could be over-estimated. At the same time, since this financial aid data was truncated, and the fact that students who did not matriculate were also offered financial aid—just not observed— probably means the influence of financial aid data was overestimated as well.

Another shortcoming of the sample was that some of the variables were not present for all five cohorts from 2011-2015. Particularly, recruiting visits were not accurately tracked prior to 2013, even though recruiting did take place. Since recruiting appears to be significant and have a positive effect on matriculation, the model for earlier cohorts might be miss-specified or at the very least omit an important variable. Moreover, high school demographics were not available for all students. Despite great efforts to obtain high school demographics—such as high school type and size—, misspellings and incomplete names made it impossible to confidently match every student's high school to a dataset from the National Center of Education Statistics (NCES).

## 5.3 Future Research

There are many steps that can be taken to improve this study. Relevant to data availability and variables, replicating this study at the aggregate university-level data would make it fully comparable to prior studies. Using aggregate data would also make findings applicable to other land-grant universities across the country. In addition, obtaining and including complete financial aid data in the analysis would likely lead to more accurate parameter estimates and better predictions. While campus visits appear to have a positive impact on matriculation, it is important to analyze other forms of recruiting as well such as mailing advertisements, online tours and high school visits.

Relevant to modeling approaches, future work should analyze matriculation decisions for transfer and international students, as these groups can comprise a large part of a universities' student body. Again however, university-wide data would be needed to have a sufficient sample size. Moreover, it might be worthwhile to follow some past studies that segregate applicants by in-state vs out of state. Some studies in the past have argued that in-state and out-of-state applicants are systematically different, and modeling them together could lead to biased estimates.

Lastly, while gradient boosting has been praised for its predictive capabilities, it appears that the logistic model performed slightly better for this sample. A possibility could be that not enough parameter-tuning took place, as interaction depth was the only tuning parameter that was varied across different specifications. Nonetheless, it is interesting that gradient boosting identified a variable not deemed significant by logistic regressions. For this reason, using two different approaches is useful because interactions that might not make intuitive sense to the researcher can be discovered. Future research should do more sensitivity analysis and perhaps

explore other machine learning methods such as extreme gradient boosting (XGBoost) or random forests.

# References


Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, *28*(25), 3083-3107.

Avery,C.,and Hoxby, C.(2004). Do and should Financial aid affect students' college choices. In C.Hoxby(Ed.), *College choices,*. Chicago, IL: University of Chicago Press.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.

Curs, B., and Singell, L. D. (2002). An analysis of the application and enrollment processes for in-state and out-of-state students at a large public university. *Economics of Education Review*, *21*(2), 111-124.

DesJardins, S.L.(2002) "An analytic strategy to assist institutional recruitment and marketing efforts." *Research in Higher education* 43.5 531-553.

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

Desjardins,S.L, Alburg, D.A andMcCall,B.P.(2006).An integrated model of application, admission, enrollment, and financial aid. *Journal of Higher Education*, 77(3), 381-429.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874

Goenner, C. F., and Pauls, K. (2006). A predictive model of inquiry to enrollment. *Research in Higher Education*, *47*(8), 935-956.

Griffith,A.,andRask,K.(2007). The influence of the U.S. News and world report collegiate rankings on the matriculation decisions of high-ability students: 1995-2004. *Economics of Education Review,* 26(2), 244-255.

Hoskin, T. (Year, Month Date Published). *Parametric and Nonparametric: Demystifying the Terms*. https://www.mayo.edu/mayo-edu-docs/center-for-translational-science-activities-documents/berd-5-6.pdf
http://info.salford-systems.com/blog/bid/337783/Why-Data-Scientists-Split-Data-into-Train-and-Test

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Johnson, I. (2008). Enrollment, Persistence and Graduation of In-State Students at a Public Research University: Does High School Matter? *Research in Higher Education*, 49(8),776-793.

Klasik, D. (2012). Common Choices: The Effect of the Common Application on Students' College Enrollment and Success. In *Annual Meeting of the Association for Education Finance and Policy, Boston, MA*

Linsenmeier, D.,Rosen, H.,andRouse,C.(2006). Financial aid packages and college enrollment decisions: An econometric case study. *Review of Economics and Statistics, 88(1),*126-145.

Monks, J. (2009). The impact of merit-based financial aid on college enrollment: A field experiment.*Economics of Education Review, 28,99-106.*

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

National Center of Education Statistics (NCES). URL: https://nces.ed.gov/

Nurnberg,P.,Schapiro,M.,Zimmerman,D(2011). Students choosing colleges: Understanding the matriculation decision at a highly selective private institution. *Economics of Education Review, 31,1-8.*

Pedro, M. O., Baker, R., Bowers, A., and Heffernan, N. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. *Educational Data Mining 2013*.

Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models*. URL: http://data.princeton.edu/wws509/notes/

StartClass. Retrieved from :http://colleges.startclass.com/compare/86-88-119/Arizona-State-University-Tempe-vs-University-of-Arizona-vs-Northern-Arizona-University. Date Accessed: March 3rd 2017

Steinberg, D. (2014,March). *Why Data Scientists Split Data Into Train and Test.* URL: http://info.salford-systems.com/blog/bid/337783/Why-Data-Scientists-Split-Data- into-Train-and-Test

Unda, A.C. (2011). *Modeling College Enrollment Decision: A Case Study of the University of Arizona* (Master's Thesis, Agriculture and Resource Economics).

US Department of Education (2015, June 4th) Retrieved from: https://ed.gov/programs/fpg/index.html?exp=0. Date Accessed: March 3rd 2017

Van der Klaauw,W.(2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review,* 43(4),1249-1287.

Weiler,W.(1996). Factors influencing the matriculation choices of high ability students. *Economics of Education Review,* 15(1),23-36.

Wolniak, G.C,Engberg M,.(2007). The Effects of High School Feeder Networks on College Enrollment. *The review of Higher Education,* 31(1),27-53.

## Appendix 1: Test of Proportions

Testing Equality of Means:

$H_o$: $P_{2011} = P_{2012} = P_{2013} = P_{2014} = P_{2015}$

| | University | | CALS | |
|---|---|---|---|---|
| Year | Matriculants | Total | Matriculants | Total |
| 2011 | 7,300 | 19,175 | 297 | 915 |
| 2012 | 7,401 | 20,251 | 394 | 1,083 |
| 2013 | 6,881 | 20,546 | 380 | 1,125 |
| 2014 | 8,023 | 24,402 | 387 | 1,012 |
| 2015 | 8,157 | 27,042 | 393 | 1,109 |
| Chi Square | 393.4842 | | 8.7707 | |
| pvalue | <.001 | | 0.0671 | |

## Appendix 2: SAT-ACT Conversion Chart (from College Board)

| SAT | ACT | SAT | ACT | SAT | ACT |
|---|---|---|---|---|---|
| 1600 | 36 | 1250 | 26 | 900 | 17 |
| 1590 | 35 | 1240 | 26 | 890 | 16 |
| 1580 | 35 | 1230 | 25 | 880 | 16 |
| 1570 | 35 | 1220 | 25 | 870 | 16 |
| 1560 | 35 | 1210 | 25 | 860 | 16 |
| 1550 | 34 | 1200 | 25 | 850 | 15 |
| 1540 | 34 | 1190 | 24 | 840 | 15 |
| 1530 | 34 | 1180 | 24 | 830 | 15 |
| 1520 | 34 | 1170 | 24 | 820 | 15 |
| 1510 | 33 | 1160 | 24 | 810 | 15 |
| 1500 | 33 | 1150 | 23 | 800 | 14 |
| 1490 | 33 | 1140 | 23 | 790 | 14 |
| 1480 | 32 | 1130 | 23 | 780 | 14 |
| 1470 | 32 | 1120 | 22 | 770 | 14 |
| 1460 | 32 | 1110 | 22 | 760 | 14 |
| 1450 | 32 | 1100 | 22 | 750 | 13 |
| 1440 | 31 | 1090 | 21 | 740 | 13 |
| 1430 | 31 | 1080 | 21 | 730 | 13 |
| 1420 | 31 | 1070 | 21 | 720 | 13 |
| 1410 | 30 | 1060 | 21 | 710 | 12 |
| 1400 | 30 | 1050 | 20 | 700 | 12 |
| 1390 | 30 | 1040 | 20 | 690 | 12 |
| 1380 | 29 | 1030 | 20 | 680 | 12 |
| 1370 | 29 | 1020 | 20 | 670 | 12 |
| 1360 | 29 | 1010 | 19 | 660 | 12 |
| 1350 | 29 | 1000 | 19 | 650 | 12 |
| 1340 | 28 | 990 | 19 | 640 | 12 |
| 1330 | 28 | 980 | 19 | 630 | 12 |
| 1320 | 28 | 970 | 18 | 620 | 11 |
| 1310 | 28 | 960 | 18 | 610 | 11 |
| 1300 | 27 | 950 | 18 | 600 | 11 |
| 1290 | 27 | 940 | 18 | 590 | 11 |
| 1280 | 27 | 930 | 17 | 580 | 11 |
| 1270 | 26 | 920 | 17 | 570 | 11 |
| 1260 | 26 | 910 | 17 | 560 | 11 |

## Appendix 3: Testing for balance in training and validation samples

Source: Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, *28*(25), 3083-3107.]

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$ **for continuous variables**

$$\frac{P_1 - P_2}{\sqrt{P_1(1-P_1) + P_2(1-P_2)}}$$ **for dummy variables (where p is proportion)**

| | N=1,554 Validation | | N=3,632 Training | | |
|---|---|---|---|---|---|
| Variable | Mean | Stdev | Mean | Stdev | Standardized Error/Difference |
| Matriculation | 0.35 | 0.48 | 0.35 | 0.48 | 0.000 |
| Northeast | 0.05 | 0.22 | 0.05 | 0.23 | 0.000 |
| Midwest | 0.05 | 0.22 | 0.06 | 0.24 | -0.031 |
| South | 0.06 | 0.25 | 0.06 | 0.24 | 0.000 |
| West | 0.82 | 0.38 | 0.81 | 0.39 | 0.018 |
| White | 0.6 | 0.49 | 0.61 | 0.49 | -0.014 |
| Black | 0.04 | 0.19 | 0.04 | 0.20 | 0.000 |
| Hispanic_Mexican | 0.07 | 0.25 | 0.07 | 0.26 | 0.000 |
| Asian | 0.02 | 0.15 | 0.02 | 0.15 | 0.000 |
| Other_Ethn | 0.18 | 0.39 | 0.17 | 0.38 | 0.019 |
| Male | 0.22 | 0.42 | 0.24 | 0.42 | -0.034 |
| Female | 0.78 | 0.42 | 0.76 | 0.42 | 0.034 |
| d_First_Generation | 0.37 | 0.48 | 0.37 | 0.48 | 0.000 |
| Median_Household_Income | 71116 | 29502.95 | 72421.87 | 30128.47 | -1.450 |
| d_High_Ability_Student | 0.02 | 0.13 | 0.02 | 0.14 | 0.000 |
| d_Honors_Admit | 0.05 | 0.22 | 0.05 | 0.22 | 0.000 |
| ACT_Max | 22.53 | 4.39 | 22.49 | 4.39 | 0.301 |
| ACT_sq | 526.64 | 200.16 | 524.94 | 201.03 | 0.280 |

| | | | | | |
|---|---|---|---|---|---|
| AP_Units | 1.94 | 2.50 | 1.95 | 2.60 | -0.130 |
| d_ABEMB | 0.03 | 0.17 | 0.03 | 0.16 | 0.000 |
| d_AGTE | 0.02 | 0.13 | 0.02 | 0.13 | 0.000 |
| d_ASC | 0.14 | 0.35 | 0.15 | 0.35 | -0.020 |
| d_ENV | 0.11 | 0.31 | 0.11 | 0.31 | 0.000 |
| d_MICR | 0.06 | 0.23 | 0.06 | 0.24 | 0.000 |
| d_NTR | 0.02 | 0.13 | 0.02 | 0.13 | 0.000 |
| d_NUSC | 0.22 | 0.42 | 0.24 | 0.42 | -0.034 |
| d_PLS | 0.01 | 0.10 | 0.01 | 0.08 | 0.000 |
| d_PRFS | 0.05 | 0.21 | 0.05 | 0.21 | 0.000 |
| d_PRRC | 0.06 | 0.23 | 0.06 | 0.25 | 0.000 |
| d_VSC | 0.28 | 0.45 | 0.26 | 0.44 | 0.032 |
| d_Other_Major | 0.01 | 0.10 | 0.01 | 0.11 | 0.000 |
| d_2011 | 0.17 | 0.38 | 0.17 | 0.38 | 0.000 |
| d_2012 | 0.21 | 0.40 | 0.21 | 0.40 | 0.000 |
| d_2013 | 0.21 | 0.41 | 0.21 | 0.41 | 0.000 |
| d_2014 | 0.19 | 0.39 | 0.19 | 0.39 | 0.000 |
| d_2015 | 0.21 | 0.41 | 0.21 | 0.41 | 0.000 |
| Days_ahead_app | 270.17 | 62.69 | 270.43 | 62.81 | -0.137 |
| Days_app_Sq | 76920.28 | 31524.23 | 77073.79 | 31647.98 | -0.160 |
| HS_Size | 1653.25 | 859.32 | 1623.19 | 850.75 | 1.157 |
| d_magnet | 0.03 | 0.17 | 0.02 | 0.15 | 0.045 |
| d_catholic | 0.09 | 0.29 | 0.1 | 0.30 | -0.024 |
| d_charter | 0.04 | 0.21 | 0.03 | 0.18 | 0.038 |
| d_public_school | 0.87 | 0.34 | 0.86 | 0.35 | 0.021 |
| d_private_school | 0.13 | 0.34 | 0.14 | 0.35 | -0.021 |
| HS_Peers | 2.11 | 2.75 | 2.11 | 2.78 | 0.000 |
| Past_Peers | 4.05 | 6.50 | 4.15 | 6.83 | -0.500 |
| Interaction_act_days_app | 6247.46 | 1865.03 | 6231.41 | 1882.36 | 0.283 |
| Campus_Tour | 0.07 | 0.26 | 0.08 | 0.27 | -0.027 |
| AZX_Visit | 0.03 | 0.17 | 0.03 | 0.17 | 0.00 |

## Appendix 4: Descriptive Statistics for predicted probabilities

| | Percentiles | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0th | 10th | 20th | 30th | 40th | 50th | 60th | 70th | 80th | 90th | 100th |
| Logistic | 0.019 | 0.130 | 0.190 | 0.242 | 0.290 | 0.335 | 0.384 | 0.456 | 0.566 | 0.723 | 0.999 |
| GBM(2) | 0.045 | 0.139 | 0.186 | 0.226 | 0.271 | 0.318 | 0.369 | 0.440 | 0.531 | 0.644 | 0.978 |

| Variable | N | Mean | Median | Std Dev |
|---|---|---|---|---|
| Logistic | 714 | 0.384 | 0.335 | 0.229 |
| GBM(2) | 714 | 0.366 | 0.318 | 0.211 |

## Appendix 5: Testing joint marginal effects (Stata)

**Non-linear combination of estimators (NLCOM)**

$$\boldsymbol{H_0: \beta_{act} + 2\beta_{act\_sq} * ACT\_Max + \beta_{interaction} * Days_{Ahead_{app}} = 0}$$
$$\boldsymbol{H_0: \beta_{days\_ahead} + 2\beta_{days\_sq} * Days\_ahead + \beta_{interaction} * ACT\_Max = 0}$$

| matriculat~n | Coef. | Std. Err. | z | P>z | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| act_MEM | -.0091366 | .0022633 | -4.04 | 0.000 | -.0135727 | -.0047006 |
| days_MEM | .0000839 | .0001454 | 0.58 | 0.564 | -.0002011 | .0003688 |

# Appendix 6a: Logistic Regression including non-test-takers

Model 1 (2011-2015, no past peers and no recruitment)

| | Model 1 (Excluding non-test takers) | | | Model 1 (Including non-test takers) | | |
|---|---|---|---|---|---|---|
| | Estimate | Wald Chi-Square | Pr > ChiSq | Estimate | Wald Chi-Square | Pr > ChiSq |
| Intercept | -1.1387 | 0.9389 | 0.3326 | 0.0522 | 0.0118 | 0.9135 |
| Northeast | -0.2559 | 1.9943 | 0.1579 | -0.3292 | 3.5341 | 0.0601 |
| Midwest | -0.4698 | 6.0404 | 0.014 | -0.5012 | 7.8046 | 0.0052 |
| South | -0.7878 | 14.4485 | 0.0001 | -0.8144 | 17.0496 | <.0001 |
| Black | -0.2563 | 1.4539 | 0.2279 | -0.1376 | 0.537 | 0.4637 |
| Hispanic_Mexican | -0.119 | 0.6232 | 0.4298 | -0.051 | 0.1371 | 0.7112 |
| Asian | 0.0827 | 0.0997 | 0.7522 | -0.0781 | 0.0997 | 0.7522 |
| Other_Ethn | -0.1316 | 1.727 | 0.1888 | -0.1141 | 1.5003 | 0.2206 |
| Female | -0.1623 | 2.8082 | 0.0938 | -0.16 | 3.2046 | 0.0734 |
| d_First_Generation | -0.0876 | 1.1162 | 0.2907 | -0.083 | 1.1635 | 0.2807 |
| Median_Household_Inc | -8.14E-06 | 33.1122 | <.0001 | -7.61E-06 | 32.9134 | <.0001 |
| d_High_Ability_Stude | -0.2942 | 0.5984 | 0.4392 | -0.2899 | 0.7117 | 0.3989 |
| d_Honors_Admit | 3.5946 | 162.721 | <.0001 | 3.4023 | 178.845 | <.0001 |
| | | 1 | | | 1 | |
| ACT_Max | 0.2372 | 8.7413 | 0.0031 | 0.1527 | 60.5263 | <.0001 |
| ACT_sq | -0.00578 | 11.5646 | 0.0007 | -0.00423 | 72.252 | <.0001 |
| AP_Units | -0.0232 | 1.8472 | 0.1741 | -0.0209 | 1.6495 | 0.199 |
| d_ABEMB | 0.1682 | 0.5792 | 0.4466 | 0.0942 | 0.2041 | 0.6514 |
| d_AGTE | 0.1952 | 0.4851 | 0.4861 | 0.3246 | 1.7493 | 0.186 |
| d_ASC | -0.1369 | 1.2413 | 0.2652 | -0.1589 | 1.9393 | 0.1637 |
| d_ENV | -0.4727 | 9.9719 | 0.0016 | -0.4403 | 9.9151 | 0.0016 |
| d_MICR | 0.0897 | 0.2864 | 0.5925 | 0.0756 | 0.2287 | 0.6325 |

| | | | | | | |
|---|---|---|---|---|---|---|
| d_NTR | 0.3421 | 1.4371 | 0.2306 | 0.353 | 1.757 | 0.185 |
| d_NUSC | 0.1075 | 1.0684 | 0.3013 | 0.0689 | 0.4972 | 0.4807 |
| d_PLS | 0.6772 | 1.9898 | 0.1584 | 0.5013 | 1.4484 | 0.2288 |
| d_PRFS | 0.4294 | 6.4621 | 0.011 | 0.3364 | 4.4967 | 0.034 |
| d_PRRC | 0.8054 | 26.3625 | <.0001 | 0.8657 | 34.109 | <.0001 |
| d_Other_Major | 0.019 | 0.0028 | 0.9578 | 0.0011 | 0 | 0.9974 |
| d_2011 | -0.0148 | 0.0145 | 0.9041 | -0.1699 | 2.1897 | 0.1389 |
| d_2012 | 0.2201 | 3.7789 | 0.0519 | 0.0615 | 0.3349 | 0.5628 |
| d_2013 | 0.0354 | 0.1031 | 0.7482 | -0.0624 | 0.3629 | 0.5469 |
| d_2014 | -0.0162 | 0.0213 | 0.884 | -0.00669 | 0.0041 | 0.949 |
| Days_ahead_app | -0.00938 | 3.9393 | 0.0472 | -0.0109 | 10.1122 | 0.0015 |
| Days_app_Sq | 0.000021 | 7.4208 | 0.0064 | 0.000021 | 9.2285 | 0.0024 |
| Interaction_act_days | -0.00007 | 0.2483 | 0.6183 | -0.00002 | 0.0744 | 0.7851 |
| HS_Size | 0.000056 | 1.3183 | 0.2509 | 0.000059 | 1.6953 | 0.1929 |
| d_catholic | 0.1135 | 0.2709 | 0.6027 | 0.0435 | 0.0445 | 0.833 |
| d_private_school | -0.0422 | 0.0457 | 0.8307 | 0.0283 | 0.0231 | 0.8792 |
| d_magnet_charter | 0.226 | 1.8992 | 0.1682 | 0.189 | 1.5776 | 0.2091 |
| HS_Peers | 0.1191 | 72.4555 | <.0001 | 0.1199 | 84.2786 | <.0001 |
| | | | | | | |
| N=3,816 | | | | N=4,597 | | |

# Appendix 6b: Logistic Regression including non-test-takers

Model 2 (2012-2015, no recruitment)

| | **Model 2** (excluding non-test takers) | | | **Model 2** (including non-test takers) | | |
|---|---|---|---|---|---|---|
| | Estimate | Wald Chi-Square | Pr > ChiSq | Estimate | Wald Chi-Square | Pr > ChiSq |
| Intercept | -0.9549 | 0.5228 | 0.4697 | -0.0313 | 0.0036 | 0.9521 |
| Northeast | -0.2427 | 1.5412 | 0.2144 | -0.3099 | 2.7036 | 0.1001 |
| Midwest | -0.5569 | 6.758 | 0.0093 | -0.625 | 9.4266 | 0.0021 |
| South | -0.792 | 12.7835 | 0.0003 | -0.7926 | 14.1951 | 0.0002 |
| Black | -0.2193 | 0.9085 | 0.3405 | -0.0907 | 0.1996 | 0.655 |
| Hispanic_Mexican | -0.0271 | 0.0277 | 0.8679 | -0.0166 | 0.0126 | 0.9108 |
| Asian | 0.0223 | 0.006 | 0.9382 | -0.1093 | 0.1592 | 0.6899 |
| Other_Ethn | -0.1213 | 1.257 | 0.2622 | -0.1063 | 1.1151 | 0.291 |
| Female | -0.1939 | 3.3573 | 0.0669 | -0.1754 | 3.2213 | 0.0727 |
| d_First_Generation | -0.0764 | 0.7087 | 0.3999 | -0.0664 | 0.6235 | 0.4298 |
| Median_Household_Inc | -7.01E-06 | 20.6647 | <.0001 | -6.93E-06 | 22.8902 | <.0001 |
| d_High_Ability_Stude | -0.3348 | 0.6346 | 0.4257 | -0.3472 | 0.8224 | 0.3645 |
| d_Honors_Admit | 3.6513 | 128.7954 | <.0001 | 3.5255 | 137.6783 | <.0001 |
| ACT_Max | 0.2183 | 5.9622 | 0.0146 | 0.1434 | 45.6007 | <.0001 |
| ACT_sq | -0.00562 | 9.0609 | 0.0026 | -0.00388 | 51.6801 | <.0001 |
| AP_Units | -0.0187 | 1.0565 | 0.304 | -0.0214 | 1.509 | 0.2193 |
| d_ABEMB | 0.1726 | 0.5364 | 0.4639 | 0.2034 | 0.8397 | 0.3595 |
| d_AGTE | -0.00965 | 0.001 | 0.9745 | 0.1623 | 0.3734 | 0.5412 |
| d_ASC | -0.2107 | 2.4209 | 0.1197 | -0.1699 | 1.8336 | 0.1757 |
| d_ENV | -0.6476 | 14.8843 | 0.0001 | -0.5976 | 14.6189 | 0.0001 |

| | | | | | | |
|---|---|---|---|---|---|---|
| d_MICR | 0.082 | 0.2017 | 0.6534 | 0.0935 | 0.2926 | 0.5886 |
| d_NTR | 0.1696 | 0.3123 | 0.5763 | 0.1926 | 0.4478 | 0.5034 |
| d_NUSC | 0.0198 | 0.0314 | 0.8593 | 0.00294 | 0.0008 | 0.9777 |
| d_PLS | 0.523 | 1.001 | 0.3171 | 0.3555 | 0.6173 | 0.4321 |
| d_PRFS | 0.2383 | 1.6478 | 0.1993 | 0.1804 | 1.0745 | 0.2999 |
| d_PRRC | 0.7728 | 20.1957 | <.0001 | 0.8537 | 27.2821 | <.0001 |
| d_Other_Major | 0.0374 | 0.0079 | 0.9291 | 0.0349 | 0.0079 | 0.9291 |
| d_2011 | 0 | . | . | 0 | . | . |
| d_2012 | 0.446 | 12.37 | 0.0004 | 0.2707 | 5.1719 | 0.023 |
| d_2013 | 0.1733 | 2.2098 | 0.1371 | 0.0636 | 0.3393 | 0.5602 |
| d_2014 | 0.0404 | 0.1278 | 0.7207 | 0.0502 | 0.2223 | 0.6373 |
| Days_ahead_app | -0.00917 | 3.0877 | 0.0789 | -0.00979 | 6.9397 | 0.0084 |
| Days_app_Sq | 0.000018 | 4.6274 | 0.0315 | 0.000019 | 6.5021 | 0.0108 |
| Interaction_act_days | -0.00003 | 0.0329 | 0.856 | -0.00003 | 0.2592 | 0.6107 |
| HS_Size | -0.00001 | 0.0575 | 0.8105 | -2.07E-06 | 0.0017 | 0.967 |
| d_catholic | 0.1189 | 0.2466 | 0.6195 | 0.0149 | 0.0043 | 0.9475 |
| d_private_school | -0.1089 | 0.2538 | 0.6144 | -0.00459 | 0.0005 | 0.982 |
| d_magnet_charter | 0.2649 | 2.3198 | 0.1277 | 0.221 | 1.9358 | 0.1641 |
| HS_Peers | 0.07 | 12.9828 | 0.0003 | 0.0734 | 16.4123 | <.0001 |
| Past_Peers | 0.0319 | 15.5831 | <.0001 | 0.0308 | 16.0996 | <.0001 |
| N=3,227 | | | | N=3,841 | | |

# Appendix 6c: Logistic Regression including non-test-takers

Model 3 (2013-2015)

| | **Model 3** (excluding non-test takers) | | | **Model 3** (including non-test takers) | | |
|---|---|---|---|---|---|---|
| Parameter | Estimate | Wald Chi-Square | Pr > ChiSq | Estimate | Wald Chi-Square | Pr > ChiSq |
| Intercept | -0.9291 | 0.358 | 0.5496 | 0.5091 | 0.6563 | 0.4179 |
| Northeast | -0.27 | 1.2789 | 0.2581 | -0.4066 | 3.1296 | 0.0769 |
| Midwest | -0.5156 | 4.0198 | 0.045 | -0.6009 | 6.1789 | 0.0129 |
| South | -0.8074 | 9.6402 | 0.0019 | -0.7985 | 10.3824 | 0.0013 |
| Black | -0.0016 | 0 | 0.9952 | 0.0348 | 0.0221 | 0.8818 |
| Hispanic_Mexican | 0.2418 | 1.6887 | 0.1938 | 0.1913 | 1.2734 | 0.2591 |
| Asian | 0.0332 | 0.0092 | 0.9236 | -0.1309 | 0.1559 | 0.693 |
| Other_Ethn | -0.0154 | 0.0146 | 0.904 | -0.0545 | 0.2131 | 0.6444 |
| Female | -0.3121 | 6.3143 | 0.012 | -0.3042 | 7.0209 | 0.0081 |
| d_First_Generation | -0.0719 | 0.4467 | 0.5039 | -0.0684 | 0.4712 | 0.4924 |
| Median_Household_Inc | -8.08E-06 | 19.0334 | <.0001 | -8.08E-06 | 21.5511 | <.0001 |
| d_High_Ability_Stude | -0.7917 | 2.1442 | 0.1431 | -0.8543 | 2.737 | 0.098 |
| d_Honors_Admit | 3.7493 | 100.4674 | <.0001 | 3.6325 | 104.5001 | <.0001 |
| ACT_Max | 0.2533 | 5.7716 | 0.0163 | 0.116 | 21.5822 | <.0001 |
| ACT_sq | -0.00631 | 8.1979 | 0.0042 | -0.00344 | 29.7048 | <.0001 |
| AP_Units | -0.0126 | 0.3586 | 0.5493 | -0.0175 | 0.7646 | 0.3819 |
| d_ABEMB | 0.1109 | 0.1786 | 0.6725 | 0.1396 | 0.3184 | 0.5726 |
| d_AGTE | 0.3179 | 0.8339 | 0.3612 | 0.3804 | 1.5886 | 0.2075 |
| d_ASC | -0.2554 | 2.463 | 0.1166 | -0.1762 | 1.3753 | 0.2409 |

| | | | | | | |
|---|---|---|---|---|---|---|
| d_ENV | -0.6471 | 10.3696 | 0.0013 | -0.586 | 9.9173 | 0.0016 |
| d_MICR | 0.0292 | 0.017 | 0.8963 | 0.0461 | 0.0479 | 0.8267 |
| d_NTR | 0.2535 | 0.4892 | 0.4843 | 0.0888 | 0.0682 | 0.794 |
| d_NUSC | 0.0146 | 0.0122 | 0.9121 | 0.0197 | 0.0253 | 0.8736 |
| d_PLS | -0.1986 | 0.0795 | 0.778 | -0.1183 | 0.0382 | 0.8451 |
| d_PRFS | 0.3473 | 2.5129 | 0.1129 | 0.3171 | 2.382 | 0.1227 |
| d_PRRC | 0.7709 | 14.2216 | 0.0002 | 0.9213 | 22.4919 | <.0001 |
| d_Other_Major | -0.1655 | 0.0965 | 0.7561 | -0.1825 | 0.1409 | 0.7074 |
| d_2011 | 0 | . | . | 0 | . | . |
| d_2012 | 0 | . | . | 0 | . | . |
| d_2013 | 0.2431 | 3.9219 | 0.0477 | 0.1167 | 1.0443 | 0.3068 |
| d_2014 | 0.0812 | 0.4761 | 0.4902 | 0.1036 | 0.8836 | 0.3472 |
| Days_ahead_app | -0.011 | 3.1791 | 0.0746 | -0.0104 | 5.3876 | 0.0203 |
| Days_app_Sq | 0.00002 | 3.828 | 0.0504 | 0.000017 | 3.3544 | 0.067 |
| Interaction_act_days | -0.00006 | 0.1102 | 0.7399 | -0.00003 | 0.1366 | 0.7117 |
| HS_Size | -0.00001 | 0.0457 | 0.8307 | 0.000015 | 0.061 | 0.8049 |
| d_catholic | 0.0031 | 0.0001 | 0.9915 | -0.1434 | 0.2795 | 0.5971 |
| d_private_school | -0.1826 | 0.4782 | 0.4893 | -0.0414 | 0.029 | 0.8647 |
| d_magnet_charter | 0.3861 | 3.868 | 0.0492 | 0.3597 | 4.0994 | 0.0429 |
| HS_Peers | 0.0179 | 0.5481 | 0.4591 | 0.0299 | 1.7685 | 0.1836 |
| Past_Peers | 0.0494 | 28.6027 | <.0001 | 0.0453 | 27.2305 | <.0001 |
| Campus_Tour | 1.45 | 107.8606 | <.0001 | 1.4399 | 115.2339 | <.0001 |
| AZX_Visit | 1.204 | 34.4143 | <.0001 | 1.2957 | 42.1747 | <.0001 |
| | | | | | | |
| N=2,493 | | | | N=2,928 | | |