Logistic Regression in the Predictive Modeling of Admitted Student Enrollment

by

Neal Ethan Logan, M.Ed.

A Dissertation

In

HIGHER EDUCATION ADMINISTRATION

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for
the Degree of

DOCTOR OF PHILOSOPHY

Approved

Michael D. Shonrock, Ph.D.
Chair

William Lan, Ph.D.

James P. Burkhalter, Ed.D.

Ralph Ferguson, Ph.D.
Acting Dean of the Graduate School

December, 2010

ACKNOWLEDGMENTS

A special thank you to Dr. Stephen DesJardins for provoking my interest in predictive modeling of student enrollment; his work provided the groundwork for this study, and his correspondence helped me emulate this study through different systems.

Table of Contents

ABSTRACT

The application of Predictive Modeling within enrollment management provides a tremendous tool for building and shaping future enrollments for institutions of higher education. Though the practice of predictive modeling is a well known application in private business practices, the use of predictive modeling in enrollment management has only recently been employed since 1990s. The independent higher education consulting firm Noel-Levitz popularized the introduction of predictive modeling as a method of providing enrollment management professionals in higher education the opportunity to forecast the possible enrolling classes of students in their institutions.

This study followed a recommended strategy for application of predictive modeling for enrollment management. Stephen DesJardins, Ph.D., from the University of Michigan published a methodology for applying predictive modeling to the process of recruiting and admitting students in an attempt to provide institutions who were actively involved in predictive modeling programs, or those who could not afford independent consulting organizations who provided predictive modeling services. The recommended method of predictive modeling as prescribed by DesJardins was adapted to an entering class of freshmen at a large, public 4-year institution of higher education in the Southwest. The class of 2009 was analyzed in order to build a model of predictive modeling which was then subsequently applied to the class of 2010. The effectiveness of both the model and the application were analyzed for effectiveness, since both of these classes have already matriculated.

List of Tables

List of Figures

CHAPTER I

INTRODUCTION

The embrace of predictive modeling applied to enrollment management for student admission practices in higher education evolves into its adolescence since its inception within the past two decades. Though a common practice within the business community, predictive modeling as applied to enrollment management has only been instigated since the mid 1990s. The introduction of this business practice began with the application of the modeling methodology by the independent consulting organization Noel-Levitz (Noel-Levitz, 2008). Noel-Levitz is a nationally recognized higher education consultation firm specializing in strategic enrollment planning for enrollment and student success.

The pursuit of predictive modeling in higher education as a means of enrollment management greatly enhances the strategic organization of the institute of higher education. Not only does the practice of predictive modeling provide a manageable structure for the administration of admission of new students and the type of student who will enroll in the institution, but it also provides a larger perspective for the institution and the annual planning of the institution. In this case, the institution of higher education can plan for enrollment logistics (such as size of classes, availability of faculty, etc.), and predict revenue generation in order to plan for budgetary

preparation. Understanding the complexity, the power, and the limitations of predictive modeling is a valuable tool for administration in higher education.

Defining Predictive Modeling

Simply, predictive modeling as a methodology, is the process of analyzing past performance of a population and applying this behavior towards a future population with similar attributes and/or seeking similar interaction. Statistically, the method of predictive modeling evolves from the family of regression analysis. It is the observed interaction of independent variables which have a positive (or negative) effect on certain dependent values that provide for this type of statistical research. Predictive research is:

> Said of an investigation whose goal is to forecast (predict, but not explain) the values of one variable by using the values of one or more other variables. Usually contrasted with explanatory research, in which the goal is to understand the cause behind relations. In other terms, the goal of predictive research is to estimate a future value of a dependent variable; in explanatory research it is to estimate the values of independent variables. (Vogt, 2005, p. 244)

In this study, it is the analysis of the discrete dependent variable (whether or not a student enrolls in an institution of higher education) which is reviewed.

Purpose of the Study

Through the course of this document, the intention of application and reliability of predictive modeling will be discussed. The application of predictive modeling begins with the evaluation of an admitted class to a large, public 4-year institution of higher education in Texas. Investigations into the characteristics of

students who select to enroll in the institution provide the groundwork for the development of the predictive model. The goal of this analysis focuses on the ability of predictive modeling's reliability in subsequent applications to subsequent populations. If this hypothesis is sustained through the analysis of subsequent administration, then the value of the predictive model will be forthcoming.

Conceptual Framework

The use of specific quantifiable tools, such as predictive modeling, provides substance and application of data sources. In this study, the data in question involves the characteristics of students enrolling into an institution of higher education. Understanding the interplay of these specific variable characteristics develops correlational relationships intrinsic to a specific student population's choice of enrollment.

Predictive modeling seeks to forecast impending results based upon the study of the values of one or more other variables. Specifically, predictive modeling is applications of predictive research wherein, "the goal of predictive research is to estimate a future value of a dependent variable" (Vogt, p. 244). Consequently, in this study, the actually forecasting will be a subsequent enrollment period (fall 2010) which has already happened. The evaluation of the predictive modeling from the characteristics of the 2009 class as compared to the 2010 class will demonstrate the effectiveness (or lack thereof) of the predictive model developed.

When evaluating the complete pool of prospective students in correlation to those students who will ultimately persist to enrollment, significant indicators within

the independent variables provided by the application of admission may become self-evident. Taking these characteristics which indicate a propensity for enrollment and applying them to the next subsequent entering class will provide a probability indication of the likelihood of enrollment in the subsequent class from the prospective student pool to the ultimate enrollment pool. Analysis of these independent variables of the student profiles will provide probability of enrollment application to be used in future enrollment populations. If the model is found effective, then future consideration about the likelihood of student enrollment can be used by the enrollment management personnel as a tool to define student populations which are more likely to enroll than those who do not show a significant probability of enrollment. For practical application, the resources and personnel of the enrollment management team can be asserted towards those students with a higher probability of enrollment, thus maximizing the return of enrollment probability for the resources invested in the particular student groups.

Assumptions

There exist assumptions which impact the effectiveness of this study. Particularly, the overall evaluation of characteristics of a particular entering class is restricted to the self-reported information which they provided at the time of application for admissions to the institution. Though predominately demographic in nature, the characteristics which make up an individual student's application profile can be interpreted differently by different individuals completing the application. The

queries are relatively one dimensional in their design, but an individual may be open to differing interpretation to the queries.

Additionally, the characteristic variables are limited only to the application set. In this case, there may be more variable characteristics among the students making application to the institution which are not a part of the application questionnaire. There are perhaps an infinite amount of characteristics variables among student applicants to college; only a minute amount of characteristics are quantified within an application for admission.

Definitions of Terms

There are a variety of definitions which are common parlance for enrollment management which are included in this study discussion:

1. Enrollment Funnel: the characterization of the student pool in the recruitment cycle; this begins with the students who are contacted by the institution as a large, potential student pool; then, as students become involved in the outreach by an institution (i.e., the students seek out information about the institution, they participate in special events, they visit the campus, etc.) they become prospective students; further through the cycle includes those students who apply to the institution, those students who are admitted to the institution, and those students who enroll in the institution. This description is characterized as a funnel since the largest pool of potential students diminishes to smaller populations as they progress through the recruitment cycle (each subsequent population of students is a smaller populations).

2. Prospect: a student who has responded to, or instigated communication and/or participation in the recruitment outreach.
3. Matriculant: an admitted student who has enrolled into the particular institution of higher education.
4. Enrollment management: the profession of higher education administrators whose role in an institution of higher education is to recruit and enroll incoming student cohorts for the institution; as a concept, enrollment management is the process by which an institution of higher education leverages personnel and resources to seek out desired potential students for incoming classes, recruit and cultivate these students to enroll in the institution, and further retain the enrolled students throughout their collegiate career.

Delimitations

This investigation into predictive modeling of enrollment to a particular institution of higher education provides invaluable research into the predictive analysis of given student enrollments. For the purposes of this study, there will be a brief analysis of two particular enrollments at a unique public, 4-year institution of higher education in Texas. The results of the study will be only applicable to this particular institution and to these particular enrollment periods.

In this study, the goal of this research will be to determine significant variables within the enrolled class of students which are observable in the potential pool of students. The variables which will be included within the study are contained within the application for admission to the institution. The study will not take into account

additional criteria which can be a part of a student admission application (e.g., letters of recommendation, resume, etc.). Efforts to maintain standardization from one year to the next will be paramount to the comparison.

Limitations

Though predictive modeling is practiced in many different environments and settings, the repetition of this study will be limited. The methodology for the process of identifying predictive trends in student enrollment may be useful in other institutions of higher education in Texas using the consistent application data. The actual data analysis, and the corresponding correlations, which are unique to the enrollment trends of this institution will remain unique to this institution in the study.

Significance of the Study

Building a predictive model for enrollment for an institution of higher education greatly enhances the strategic prioritization and resource management within the institution. By developing a model of potential enrollment with some accuracy, an institution of higher education can implement earlier models of budgetary planning, resource allocation, and personnel appointments. Though it is practically impossible to predict enrollments with 100% accuracy, predictive analysis for enrollment trends ultimately promotes a better preparedness within the institution of higher education.

Organization of the Study

The discussion of developing a predictive model for admitted student enrollment in this study builds upon the study of historical enrollment patterns and student cohorts. For this study, a historical cohort of admitted student enrollment will be evaluated in order to develop a model of predictive analysis which will be used on a subsequent student population. A review of the relevant literature will provide conceptual framework for the idea of qualities of student college choice; this foundation will provide the research with grounding in the contributing factors influencing college student choice.

The method of building a correlational study around the conceptual framework of predictive modeling will be discussed in the methodology section of this presentation. A discussion of the theoretical and operational methods used to develop a predictive model of admitted student behavior will be discussed and presented in an operational format. Once the methodology of the study is presented, the actual data analysis of the model will be revealed and discussed in the findings portion of the study. Finally, conclusions will be drawn from the study whereby the effectiveness of the predictive model will be discussed and the potential for future application will be considered.

CHAPTER II

ANALYSIS

Introduction

Research in the application of predictive modeling for enrollment management eludes most of the common literature on this topic. More often than not, the treatment of predictive modeling in the literature merely coincides with the application of the process; the handiwork of the tool. Among tools of data analysis, predictive modeling is a powerful application towards prediction of future enrollment of students. In the review of pertinent literature, this study will review both the applications of data analysis for the purposes of predicting future enrollments and also review the underlying theories which have been attributed to the process of a student's choice in attending a college or university.

Student College Choice

In the advent of the 21st century, more and more students are entering into higher education to continue their education and to further their future career goals. Balanced within the number of students seeking entrance into higher education are the colleges and universities who are competitively recruiting these students to their institutions and who seek to find those students with the highest quality profile which reflects upon their desired entering student body. These students are the ultimate

consumers who find themselves trying to find the "best fit" for their postsecondary educational needs and who are courted by these various colleges across the country.

In the present American landscape of higher education, every stake holder (students, parents/families, policy makers, etc.) generally believe that everyone in our society should partake in and benefits from postsecondary educational opportunities. Invariably, many researchers supplement this assumption that post secondary education provides directly proportional higher salaries, more productive working lives, more career opportunities, and an increase in the quality of life for the recipient (Bowen, 1977; Leslie & Brinkman, 1988; Pascarella & Terenzini, 1991). Postsecondary education presents more opportunity and greater return on investment to those students who choose to pursue advanced education.

Historical Setting

In the beginning of the 20th century leading up to the 1940s, college bound high school graduates averaged approximately 20 percent of all high school graduates (Kinzie, Palmer, Hayek, Hossler, Jacob, & Cummings, 2004). These students were primarily students from affluent families whose parents, more than likely, had also participated in higher education. The path towards higher education was marked primarily by means of access and familiarity with institutions of higher education. Generally, students had little interest in researching possible institutions of higher education and were essentially comfortable with the reliance on vagueness of institutional reputation when deciding upon which college to attend (Holland, 1958).

The esoteric nature of collegiate matriculation would change in America, however, after the conclusion of World War II.

After the second great global conflict, perhaps the single most important event in American higher education evolved with the institution of the GI Bill (Servicemen's Readjustment Act of 1944). The toil of war left many institutions of higher education struggling with the lack of enrollment and resources which were consumed in the investment of America in World War II. With the hope of both reintegrating servicemen back into society and with the need for federal investment to revitalize American higher education, this historical legislation brokered a successful fusion of two important constituents in American society. The GI Bill provided just that; providing federally funded tuition and books to veterans which prompted significant educational development to the work force while bolstering college enrollments, consequently opening access to college and universities to a level yet unseen in American higher education (Rudolph, 1990). The prism of college choice began to change as prospective students began to see the possibilities of the differences of institutions of higher education; no longer were students simply going to local colleges or those colleges of note within their own milieu. Postwar higher education was marked with increasing investment and participation in higher education, spawning a cultural shift in the college-going culture of America.

The changing landscape of student interest in higher education and the expanding enrollment of these students brought change to the work of the college admissions personnel. Consequently, the rise in the professional admissions personnel

prompted the inclusion of "and Admissions Officers" to the already established American Association of Collegiate Registrars in 1949 (American Association of Collegiate Registrars and Admissions Officers, 2002). Increasing student participation in the college selection process prompted an equal expansion of admissions personnel who worked with the increasing pool of candidates as well as recruited selectively those students who they wished to see enter their institutions. Admissions personnel, who had also already been organized in the National Association for College Admission Counseling in 1937, contributed to the evolution and development of the standardization of collegiate admissions in association with such groups as the College Board and the American College Testing (ACT) organizations. Together, these groups built the foundation of the collegiate admissions process, and by the mid 1950s collegiate entrance requirements included such standards as college entrance exams, high school transcript submission, high school curriculum preferences, recommendations and interviews, and high school class rank as components of college entrance submission (Beale, 1970).

The postwar expansion of higher education investment and enrollment experienced unique shaping in the decades to follow. Access, as well as equity of access, drove the corresponding enrollment growth patterns within higher education. The largest group of students entering into the higher education market arose during this period:

> The postwar "baby boom" added to the tidal wave in undergraduate student enrollment. Although the front edge of the "baby boom" generation did not turn 18 until 1964, this group, along with many high school graduates who chose college instead

> of going directly to work, greatly increased the student population in the 1960s. (Kinzie, Palmer, Hayek, Hossler, Jacob, & Cummings, 2004, p. 15)

Student enrollments benefited from interest in education at unprecedented levels as they built upon the foundation of an ever-increasing college-going heritage from the previous generation. Additionally, access to higher education for these students entering into colleges and universities expanded at an equal pace.

Federal legislation paved the way of access to upcoming students in higher education. Under the administration of President Lyndon B. Johnson, the Higher Education Act of 1965 opened to doors of collegiate education to economically disadvantaged students through federal financial aid programs (Brubacher & Rudy, 1997); this legislation was expanded upon in 1968 reauthorization of the act to include programs of opportunity for all students regardless of their racial/ethnic origin or economic situation (Wolanin, 1996). Additional federal involvement in higher education policy continued to expand access. Affirmative action programs and the Title IX subsection of the 1972 Federal Education Amendments dissolved barriers (especially relating to gender) to college access for the American student (Miller, 1999). College choice opportunities continued to develop with the landmark case *Adams v. Richardson* (480 F.2nd 1159 [DC Cir. 1973]), requiring states to respond to discriminatory practices against students of color by enforcing desegregation, increasing the diversity of students, faculty, and staff, and enhancing minority student recruitment and retention.

The widening access towards enrollment of students in the United States promoted increasing pools of prospective students for colleges and universities. The economics of availability, however, were somewhat mitigated by concerns about the academic preparedness of these potential new students. Increasing availability of access to students (and the expectations of society for inclusionary efforts) directly conflicted with the interests of those who felt that increasing access decreased the quality of the applicant, and in turn, decreased the quality of the college or university (Kerr, 1990). These two elements contributed to the expansion and integration of admissions management in institutions of higher education. Competition for students became an effort of attracting quality of academic potential and enrichment of the institutional profile. Corporate-style marketing began to emerge as institutions enhanced their archaic direct mail protocols with professional media development (Duffy & Goldberg, 1998). The era of the college "view book" emerged.

The student perspective towards future enrollment into institutions of higher education provoked a different response. Now as a prospective student, the economic model was supply-side heavy with respect to the higher education market. Selectivity favored the institutions of higher education; so students began to demand more for their investment. The late 1970s marked the change in the college-going culture of prospective students. These students viewed themselves as consumers who invested their monies in higher education only after making sure that they would get the best return on their investment (Chapman, 1978). Understandably, parents of prospective students remained the most influential contributors to student choice (Welki &

Novratil, 1987), specifically in terms of affordability and location, along with counselors, teachers, peers, friends, and relatives who all had influence on the prospective students choice.

The 1980s and 1990s experienced tumultuous extremes in student diversification, rising costs of higher education, and subsequently declining federal support for higher education. Choice and access, were buffeted by the Supreme Court's ruling in *Regents of the University of California v. Bakke*, 438 U.S. 265 (1978) allowing for racial consideration as an admission criteria for prospective students to the case of *Hopwood v. Texas*, 78 F.3d 932 (5th Cir.), cert. denied, 518 U.S. 1033 (1996), where race was not considered a contributing factor for admission consideration in the district court, though overturned by the Supreme Court appeal. College and university admissions administration dealt with the changing scope of underrepresented students whereby race/ethnicity no longer provided a compelling interest in collegiate admissions. Shrinking federal support for higher education in the 1990s meant that students with low socio-economic backgrounds (arguably more predominately affecting underrepresented students) increased the reliance on loans and decreased the likelihood of enrollment of these students with low socio-economic backgrounds (McPherson & Shapiro, 1993). College choice for students began its relative decline of possibility while fortifying the consumerist mindset of best fit in choice for the investment of higher education dollars.

College Choice and Enrollment Management

Modern enrollment management strategy, adapting from the historical trends of student choice, both tempers the environment of potential college enrollment and bolsters a greater commitment to research and design for the enrollment process from beginning to end. With the ever-increasing downturn in federal and state funding of higher education in the public higher education market, enrollment managers in colleges and universities have been focused upon developing more competent and productive models for enrolling students as well as anticipating what the student enrollment will look like once it comes to fruition; "they have wanted to more effectively plan and forecast their enrollment, and to more effectively influence the enrollment decision-making process of prospective students" (Paulsen, 1990, p. 6). Paulsen (1990) provides a context for reviewing the study of the entire prospective student body as the macro-level of college choice behavior which includes two distinct parts for consideration in the prospective enrolling body: behavior in relation to environmental factors (outside of the institution's control) and institutional characteristics (within the locus of control for institutions).

Juxtaposed upon this macro-level of consideration of gross student enrollment choice behavior are fundamental theoretical frameworks which seek to derive choice modeling. Paulsen (1990) identifies three particular, unique theoretical foundations to frame his macro-level choice behavior system: the sociology perspective, the economic perspective, and the psychology perspective. A more modern theoretical

perspective, the fourth consideration, derives from DesJardins's review of student choice literature.

The sociological perspective focuses upon the status attainment process in the college choice behavior. Paulsen (1990) argues that perhaps the most contributive research to early stage choice behavior by students occurs within this process, with students forming early educational aspirations about attending college. Indicative as contributing factors for status attainment processing, student characteristics which are important considerations within the status attainment model include characteristics of a student's family, their high school background, and the student's academic ability.

Economic behavioral theory which influences college choice behavior is characterized as investment decision-making behavior (Jackson, 1978). The economic paradigm views the college choice behavior in terms of and investment scenario, with weight on the return on investment. Student background and academic ability are used throughout this model as constant factors which interact with the environmental and institutional characteristics of the macro-level choice modeling.

The psychological framework for college choice originates in the seminal work of the psychologists who research students in higher education. Alexander Astin (1965) provides the foundation of college choice as:

> The characteristics of the students enrolled by an institution are highly related to measures of the psychological environment or 'climate' of the institution. If, as these findings suggest, the college environment is determined to a large extend by the kinds of students at the institution. (Astin, 1965)

This theoretical construct surmises that the tendency for students to attend institutions of higher education with similar characteristic like themselves is a natural process of

attenuation of student choice behavior. The corollary for institutions seeking to change these characteristics, and influence the prospective student class characteristics of the future, is to actively recruit (and enroll) those students who embody the characteristics of the institution's desired student body.

DesJardins (2002) posits college student choice behavior incorporates both the college students' individual characteristics interacting with their preferences of the institutional characteristics of colleges and universities which they are considering (Fuller, Manski, & Wise, 1982; Manski & Wise, 1983). He further extrapolates macro-level application of student college choice behavior as a function of the characteristics of a population of potential students in relation to the characteristics of relevant existing institutions of higher education (Hoenack & Wieler, 1979). Building upon these foundations, DesJardins provides a college choice model based upon three broad stages. The first stage involves the formation of college aspirations; a period of time when a student becomes aware of higher education and the likelihood of pursuing advanced education which typically lasts from early childhood through early high school. The second stage (known as the "choice" stage), involves identification of colleges and/or universities, selection of a pool of potential institutions, and application to this pool. The "choice" stage involves the historical, traditional student age group of juniors and seniors in high school (or later for nontraditional students). The third stage involves actual admission to these chosen institutions (some or all) of higher education and the eventual enrollment into one of these possible institutions.

The macro-level college choice behavior, evident in these college choice theoretical foundations, attempt to characterize the choice and enrollment of students in aggregate. In enrollment management, these tendencies provide predictive characteristics when reviewing potential student populations. Linear regression, "is the most common statistical procedure used to analyze data on the behavior of student groups" (Paulsen, 1990, p. 8), and thus can be effectively used in the process of prediction of future enrollment based upon a given population's characteristics. Moreover, these analyses of student populations can be applied to the environment and/or institutional characteristics of a given institution, and influence future policy development, environmental scanning, planning and forecasting of future enrollment at a specific institution (Hossler, 1984).

Analytical Strategies for Enrollment Management

An ever-increasing challenge and expectation of enrollment management in institutions of higher education involves the ability to predict the efficacy of effort by the enrollment management personnel and processes on the future enrollment of the host institution. Fundamentally, enrollment management as a profession, provides not only the work to recruit and admit desirable students to the host institution, but also predict the future enrollment of these students with increasing accuracy as the institution approaches the matriculation date of a given term. The benefits of predicting, with some accuracy, the future enrollment of students in a given future term for an institution of higher education are numerous. Primarily, tuition revenue forecasting provides important budgetary direction for future planning of the

institution. Specifically, with the financial burden of institutions moving away from federal and state support towards tuition revenue as the primary revenue source:

> [O]perating costs have escalated and public-sector financial support has flattened. As a result, many colleges and universities have had to sharply increase tuition and fees and look for ways to control costs to avoid financial disaster. (Council for Aid to Education, 1997)

Additionally, institutions can prepare for student enrollments in such cases as providing student housing estimates for occupancy, scheduling classes and arranging faculty load, and providing adequate student support for the incoming class.

The increased reliance upon tuition revenue alone invigorates the needs of institutional enrollment management strategy in recruiting and enrolling students; but the climate and nature of modern enrollment management also includes the development of the characteristics of the incoming classes. Enrollment management additionally operates with "the pressure to enroll more high-ability students, and the desire to have a diverse student body effective recruitment and enrollment of students, is an even more important function than it was a decade ago" (DesJardins, 2002, p. 531). To meet the multiple target variables of a desirable class, enrollment managers seek to maximize effect, while minimizing the costs associated with recruiting a class which ideally would be high-ability and diverse. There are a multitude of tools and analysis which are used and combined in order to facilitate these desired results.

Applications of Enrollment Management Data Analysis

In increasingly competitive markets for prospective students, institutions of higher education include a variety of recruitment tools and techniques in order to

appeal to students. Among these tools, methodologies for reviewing the efficacy of recruitment tactics provide two insightful products. First, analyzing the data from year to year among college bound students within the context of the methods of recruitment for a particular institution allows for a measure of the effectiveness of each technique, program, instance, etc. By analyzing effectiveness of a given product, the return on the resource investment can be measured. Though not completely capable of cross comparison between one application and the next, some activities will describe a higher yield of success than others. Secondly, applying this analysis forward toward future investment of resources, the recruitment of future students can become more focused on student events, programs, recruitment techniques, etc., which provide for more effective recruitment when compared to previous year's results.

Historical Trend Data

Perhaps the first method of critical review for data analysis for enrollment management invokes the review of the historical trends of enrollment at a particular institution of higher education. The strength of the historical trend model "…provides a solid foundation for building effective marketing and recruitment strategies" (Noel-Levitz, 2007). The establishment of the historical record for enrollment of students can be parceled into a variety of subgroups and categorical descriptors. This application ranges from the simple consideration (e.g., total enrollment from one year to the next), to the complex contagion of variables (e.g., geographic market, academic performance, college entrance exam scores, financial status, etc.). Reviewing historical trend data promotes the most solid data to the point of prediction. Future

forward assumptions can only be made as a basis of historical trend indications. Successive years of increasing enrollment, for example, do not necessarily guarantee the next years continuation of enrollment growth.

Operational Research

Where historical trend data analysis tracks the trends of enrollment across the students and their specific characteristics, analysis of operational effectiveness measures the application of enrollment management through its various forms and the success of each of these applications. This method of data analysis examines each of the elements of a college admissions office and the recruitment, marketing, and processing of the student's participation in a recruitment cycle. By analyzing these elements and isolating them independently, the effectiveness of their application may be measured. The benefits of operational research provide information where "an institution is able to focus its limited resources on those strategies that are most likely to generate the desired enrollees" (Noel-Levitz, 2007). In this method of analysis, an institution isolates particular facets of recruiting student interaction and gauges the impact of said interaction through specific indicators. In most cases, these indicators are typically the pivotal points of a student enrollment cycle (i.e., making application, being admitted and enrolling in the institution).

Cost-Benefit Analysis

In an encompassing definition, cost-benefit analysis provides the total cost of recruitment and matriculation of prospective students. Holistically, this measure of

cost per student recruited provides a framework for analysis of total gross expenditures for return of enrollment. Exponentially, each aspect of the student recruitment process can be dissected and reduced to a cost per student as they progress through the enrollment funnel. Through cost-benefit analysis, "…enrollment managers link budget data to activities such as direct mail and travel to arrive at a specific cost-per-matriculant" (Noel-Levitz, 2007). Arguably, the evaluation of each aspect of recruitment can be itemized for comparison; those programs with the highest cost may not necessarily have the most success in recruiting students and can be marginalized, reduced, or removed.

Market Research

Market research seeks to derive information directly from the population with whom there exists the desire to impact their decision making process. In the case of the enrollment manager, the most important population is that of the prospective students who may enroll into the institution in the future. Secondarily, market research seeks to ascertain information about those populations who influence these prospective students in their daily lives, whether they are their parents, their school counselors, or similar contributing constituents (Noel-Levitz, 2007). Typically, market research in the area of marketing to prospective students includes these students themselves, but also other categories of populations, such as: lost inquiries, lost applicants, incomplete applicants, lost admitted students, prospective employers, and graduate school programs. The most typical approach to learn more about these populations includes the practices of surveys and focus groups of the populations in

question. A more passive approach to market research includes the use of environmental scanning and competition studies, both analyzing the external factors which interrelate to the prospective student and their decision making processes.

Secondary Data Resources

In addition to the different methodologies used to provide information about prospective students and the effectiveness of enrollment management strategies, there are additional data sources provided to the enrollment manager through public and private institutions which explain and predict the future student's environment and attitude towards higher education. Examples of these resources include, but are not limited to, the state/national demographics for college going students, the income distribution and resource potential of geographic markets, population growth, and characteristics of college-going students through such agencies as the College Board, ACT, and the NRCCUA (National Research Center for College and University Admissions).

Student Price Sensitivities

With the cost of higher education looming as a significant investment for prospective students and their families, the "willingness to pay" provides a point of consideration in institutional research. A critical element in every prospective student's choice of attending an institution of higher education will be the actual cost of attending, coupled with the amount of financial aid and/or scholarships which will offset the cost of attendance. This research provides institutions with the leveraging

capacity to determine exactly at what price point a student is affected in their decision making process. More specifically, at what level of financial aid, scholarships, and/or tuition discounting does the student find the institution to be a good value?

Predictive Modeling in Enrollment Management

Predictive modeling, in essence, is the analysis of past performance used to predict the future application of the same methodology. In enrollment management, "predictive modeling is frequently used to identify students most likely to apply or to enroll in a college or university so that admissions staff can concentrate their attention on these 'hot prospects' in order to enroll more students" (Gose, 1999). There are a variety of student populations that prediction can be applied: the prospective students may be categorized by likelihood of making application to the institution of higher education based upon geographic distance from the said institution, the student applicant's characteristics can predict the likelihood that they will enroll, and the enrolled student's persistence and retention can be predicted by their academic performance within their first term as a student in higher education.

To enrollment managers, this second example of the admitted student's likelihood of enrollment is particularly important and the focus of significant investment. These "hot prospects" who have made application to an institution have demonstrated a level of interest and investment to the institution; in the enrollment management profession, this level of investment is second only to the actual matriculation of the students to the institution of higher education. Though this action in itself is significant, today's college student may have applied to more than one

institution in the process of making their college choice. To qualify these students further, enrollment management professionals seek to measure and influence the level of investment and likelihood of enrollment of these students through communication.

The purpose of the continued communication to admitted students involves assessing levels of interest (Kahler, 2008). Though a student applies to an institution of higher education, the act is not necessarily a declaration of commitment. The key component of the assessment of interest involves the type of communication put forward for a particular type of student. There are a variety of measures and classifications for levels of interest when qualifying admitted students as potential matriculants to the institution. With limitations of time and resources, institutions seeking to capitalize on enrolling the most desirable students and student classes:

> Rather than focusing on students who are very likely to enroll, enrollment consultants often direct institutions to focus on students who are at the margin with respect to enrollment. These 'fence sitters' (as some analysts call them) are students who may be convinced to matriculate to the client's institution. In collaboration with enrollment management professionals, consulting firms use statistical techniques to estimate each admitted student's probability of enrollment. (DesJardins, 2002, p. 532)

Flexibility in application of resources maximizes investment on behalf of the institution of higher education; the use of predictive modeling to qualify the levels of interest of students in correlation to the likelihood that these students will enroll provides a powerful tool for the enrollment manager to direct limited time, resources, and personnel more effectively.

Predictive modeling employs the characteristics of student interaction, descriptor values, and demographics in a statistical model in order to forecast the

likelihood of their actions (for the purposes of this study, their likelihood of enrollment). Common data sets of these characteristics for student populations are used for both consistency and generalizability. These attributed characteristics "whether relating to socio-economic status, geographic setting, academic achievement, attendance at recruitment events, communications to the school, or even time and method by which they entered the enrollment funnel – all these variables may contribute to the model" (Kahler, 2008, p. 147). This application provides stratification of students based upon the probability of enrollment, allowing enrollment management professionals the ability to categorize students from "definitely will enroll" to "unlikely to enroll". The "fence sitters", or the middle range categorization of students, however, represent students who "may enroll" and are of specific interest to the enrollment management professional.

The "fence sitters" provoke a call to action for enrollment management professionals and are often the most important population of students in a recruitment cycle and the subject of most enrollment management consultant's direction to their clientele (DesJardins, 2002). These students are operationally the students who have the most susceptibility to encouragement from the recruitment efforts of an institution of higher education. Those students who are classified as "definitely will enroll" should not be ignored in the continuation of recruitment efforts; however, their declaration of intent will be sustained very economically. The "fence sitters" represent a group of students who have not yet committed to the recruiting institution; these students represent a potential pool of admitted students who may yet be

convinced to enroll at the institution. The enrollment management personnel today regularly regard these potential students with the upmost of consideration and application of resources in order to try to influence their decision to enroll in the institution. Those students who definitely will enroll provide a strong reliability of contribution to the incoming class. The potential student pool provides the greatest amount of additional gain since their likelihood to enroll is still in the balance. Enrollment management theory addresses these students as the students where the greatest effort of resource investment in a given recruitment cycle is best attenuated. By focusing on the investment effort on these potential students, the gain of the entire entering cohort may be bolstered in number.

CHAPTER III

METHODOLOGY

Predictive modeling in enrollment management promotes an analytical and systematic approach to the art of forecasting the future enrollment of an incoming class of students to a particular institution of higher education. To this end, the institution builds upon the past performance and trends of known student behavior and applies this data forward onto the future student population. There are a variety of methods of statistical application to evaluate the characteristics of a population of an enrolled class into a given term of an institution higher education. This study will endeavor to provide analysis of the successful student matriculant and apply this profile forward onto a future incoming class. By evaluation and application of a successful predictive model, a significant tool for enrollment management professionals can be developed and refined for subsequent applications.

Research Questions

Predictive modeling in enrollment management analyzes the population of historical student characteristics in an effort to apply the logic of the actions of these historical student groups (and the choices that they have made, e.g., admitted students who enroll in an institution of higher education) to a future student group by comparing the same characteristics consistently in the framework of the like choice sets. For the purposes of this study, the characteristics of a historical student population of admitted students who enrolled into the sample institution of higher

education will be applied to the future enrollment population to generate a probability of enrollment for the future class. To this end, the following questions will be evaluated:

1. Of the characteristics of an incoming admitted student class considered, what is the predictive accuracy of enrollment for these students? Specifically, what is the effect of the independent variables (student characteristics) of the admitted student class which influence enrollment (dependent variable)?
2. Does this prediction of enrollment demonstrate statistical validity?
3. Does the predictive nature of an admitted student class enrollment provide evidence of prediction for a subsequent future admitted student class enrollment? Specifically, is the model an effective predictor of a subsequent admitted class's enrollment?

Rationale and Data Analysis

In applying predictive modeling upon a prospective admitted student class of an institution of higher education using the model of enrollment of a previous admitted student class, forecasting models (regression models) provide the best method of likeliness of choice among these student populations. Specifically, logistic regression analysis typically is used to provide probability of choice among the students within the models application. A historic group of admitted and enrolled students provide and explanation of the choice of enrollment which can be used to predict the enrollment behavior of a subsequent class of students, where "[t]he results can be used to explain enrollment behavior by examining how the independent effects of the

included regressors affect the probability of enrollment" (DesJardins, 2002, p. 538). Building upon the revealed information, the choice of student enrollment (of each admitted student) can be predicted by segmenting groups of students into their respective propensity for enrollment.

To address the likelihood of enrollment of admitted students to an institution of higher education, prediction analysis fundamentally begins with the concept of regression analysis. In cases where prediction is a function of the question of impact (the dependent variable) by an independent variable, linear regression provides the fundamental framework for forecasting an outcome. That is, "for each unit increase in the independent variable, we would expect to see a change of some fixed number of units in the dependent variable" (Dey & Astin, 1993, p. 571). Linear regression does, however, have limitations of universal application. Nonetheless, the "assumption of a linear model is usually the starting point for analysis," because of its ease of use and comprehension, and because it can be applied to the study of nonlinear relationships (Hanushek & Jackson, 1977, p. 25).

The challenge with applying linear regression models to the study of the sum total of student characteristics of a admitted student population (the entirety of the independent variables) and their interaction on the choice for enrollment (a dichotomous dependent variable, i.e., either enrolled or not) centers around the assumption that the dependent variable exists on a continuous scale. In the case of student enrollment as a dichotomous dependent variable, there is only the choice of enrolled or not enrolled, there is no continuous variability of the dependent variable

(e.g., there is no “almost enrolled”). According to Dey and Astin (1993), these linear models can sometimes predict values for dichotomous variables which have no meaning (such as negative probability), and when dealing with probabilities the relationship is not always truly linear. Changes in the independent variables have a higher likelihood of influence on the probability of something occurring in the middle of the range of probabilities than in the extreme ends of the range.

In order to better understand the relationship between admitted student characteristics and their choices for enrollment, a more complex methodology of regression analysis provides a better understanding of the observed behavior. Due to the nature of the outcome of the dependent variable, that of binary or dichotomous, which is not on a continuous scale, the more appropriate regression analysis application is that of a logistic regression model (Hosmer & Lemeshow, 2000). More expressly, the binary/dichotomous nature of the dependent variable is a curvilinear relationship, not linear at all and can best be graphically described in the following figure (Cabrera, 1994):

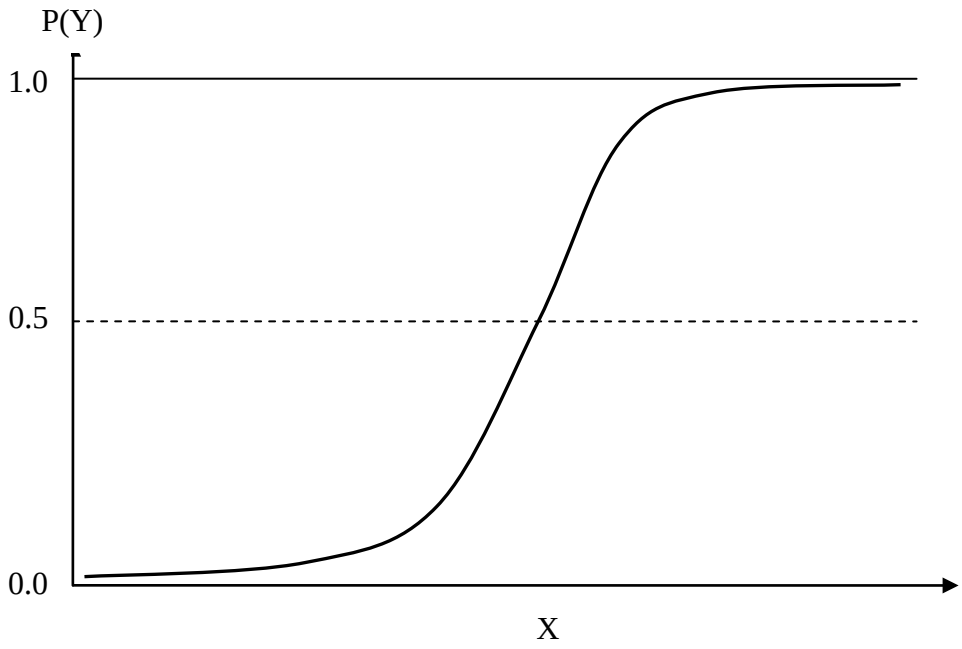

Figure 1: The Logistic Distribution

In Figure 1, the probability of the dichotomous variable (Y) enrollment criterion (for the purpose of this study) follows the default threshold of either ($P > 0.5$) enrolling or ($P < 0.5$) not enrolling. The curvilinear nature of the regression model demonstrates that there is greater influence of the independent variables on the middle ranges on probability, than on either ends of the range. The logistic regression model dismisses the continuous variability of the linear model (i.e., which would be represented by a straight line across the probability of the outcome) and provides a greater emphasis on the significance of the threshold probability of the outcome. That is, the probability of an event outcome (a dichotomous variable) is either going to occur or not (simplified, a 50% chance of occurrence).

Logistic Regression

For the purpose of this study, the methodology for application of the logistic regression model for prediction of enrollment of future student populations follows the methods prescribed by Stephen L. DesJardins, from the University of Michigan. DesJardins provides the framework for evaluation of the potential predictive modeling of enrollment behaviors from one entering class to the next by comparing a completed enrollment cycle to a subsequent enrollment cycle using the consistent measures of independent variables between each population. His prescribed methodology and testing will serve as a model for this study in an attempt to replicate these methods in case of an independent student population and institution of higher education.

The DesJardins method incorporates the logistic regression model, providing a probability model for admitted student enrollment across the various independent characteristics which make up the incoming student class profile. The logistic regression model used by this study is described by

$$\log \frac{P}{1 - P_i} = \alpha + \beta_i X_i + \delta_i Y_i + \gamma_i Z_i + \varepsilon_i$$

where P_i is the probability that student *i* will choose to enroll in the study institution; X_i is a vector of the personal and demographic characteristics of the student population (e.g., socioeconomic background, academic ability, etc.); Y_i is the vector of prior educational characteristics, college intentions, and preferences; Z_i is an institutional variable indicating the historical enroll/applicant ratio for the student populations high schools; $\alpha, \beta_i, \delta_i$ and γ_i are estimated coefficients; and ε_i represents a

random error term that is logistically distributed. The dependent variable is the logarithm of the odds that a particular student will enroll in the institution in the study (DesJardins, 2002).

Maximum Likelihood

The formula for this replicated study incorporates the elements of the logistic regression model and includes a method of estimation for the unknown parameters of the independent variables. In linear regression models, the *least squares* method minimizes the sum of squared deviations from the observed variability of the dependent variable. In logistic regression, however, "when the method of *least squares* is applied to a model with a dichotomous outcome the estimators no longer have the same properties" (Hosmer & Lemeshow, 2000, p. 8). Instead of the *least squares* model, the approach to yield the best estimators for this method involves the *maximum likelihood* model. Generally, the *maximum likelihood* approach yields values of the unknown parameters which maximize the probability of achieving the observed set of values (Hosmer & Lemeshow, 2000). For the purpose of study, the *maximum likelihood* is the *log likelihood* value with the best parameters of prediction in the design of the analysis. The DesJardins method in this study relies on the operation of *maximum likelihood* estimation.

Application of the Model

The application of the DesJardins model begins with the assessment of the logistic regression of the historical data set of admitted student enrollment. For the

purpose of this study, a previous term enrollment of students will serve as the historical data set. The estimates that are determined from the application of this model on the historical data set will then be used to calculate the probability of the future enrollment set of admitted students and their enrollment decisions. This practice is referred to as "scoring" the data set of students by their probability of enrollment (DesJardins, 2002). Once scored, these student groups are segmented into roughly equal number of students (the recommended usage being deciles of student groups). The further application of this process provides the enrollment management professional with knowledge of the probability of enrollment of students within these segments, with the intent of influencing the "fence sitters" with more application of time and resources to encourage their further development towards matriculation.

In order to avoid the possibility of avoid bias and overly optimistic prediction from one enrollment class to the next, the research contends with the problem of the accuracy of the probability development. Inherently, the use of "…the same data to test the predictive accuracy of your model that you use to fit the model, it biases your results (SAS Institute, Inc., 1995, p. 36). DesJardins recommends the adoption of a strategy to minimize this bias by randomly splitting the historical data set into a "developmental" sample and a "validation" sample. In this process the developmental sample is used to build the logistic regression and the validation sample (sometimes known as the "holdout" sample) is used to test the model. The tested model can then be used to apply towards a future model for the predictive application.

Model Fitness

In the discussion above, the process of building a developmental sample and testing the model on a validation sample requires a testing of the model's fitness before the model can be used on subsequent populations. Typical fitness models are build around the logistic regression process involves a 2x2 table of "hits and misses of a prediction rule (Greene, 1993, p. 651). The common threshold for probability prediction centers around the 0.5 probability (and above for positive correlation), which is to say a 50% chance (or greater) of enrolling a student based upon the model. DesJardins argues that this method may not be appropriate for the purposes of applying prediction on incoming student enrollment due the inherent variability of the populations themselves. If the distribution of admitted student enrollment is unbalanced (with a large number of students who do enroll, or vice versa), then there exists a chance of error assigned at the typical threshold value. He states that "[g]enerally, there are two types of errors to consider when assigning a cutoff score: the incorrect classification of enrollees and incorrectly classifying non-enrollees" (DesJardins, 2002, p. 540). To address this concern, a logical response would be to establish a different, more conservative, threshold value. This decision will however overbalance the model with more students predicted to enroll who, in actuality, did not.

The DesJardins methodology which will be used in this study, actually employs multiple models of fitness testing in analyzing the model. He recommends that the sensitivity of the model fitness can be used in the deciles segmentation that is

already in place for the variability coefficients. The actual enrollment vs. predicted enrollment testing can be arranged across these deciles and is a popular methodology used in logistic regression classification (Hosmer & Lemeshow, 2000). This is accomplished using HL goodness-of-fit statistic which tests the null hypothesis that the model fits the data (thus, a significant HL test indicates that the model does not fit the data) (DesJardins, 2002). Additionally, the accompanying Brier score (Brier, 1950), a score which is a unitless index of predictive accuracy, is further used to illuminate the predictive accuracy of different models and provides further justification of the fitness of the model.

Context of the Study and Data Sources

The study of the predictive modeling of enrolling students in this design will consider two subsequent populations of admitted students to a large, public 4-year institution in Texas. For this study, the enrollment predictive modeling will be developed from an entering class of freshmen students from the fall term of 2008. The development of the model from the fall class of 2008 will then be applied to the entering class of freshmen for the fall of 2009. Since both of these entering classes have already matriculated, the further assessment of the model can be gauged by the actual vs. probable enrollment of the class of 2009. The institution is a traditional public 4-year institution with a large (majority) of the entering class in the fall term being freshmen students. Discriminating factors from this sample towards generalizability of the model include the geographic region and state in which the enrolling classes are measured. Additionally, Texas is currently a growing population

of college-age student; a climate which is not necessarily replicated in all other states within the United States.

These two entering cohorts of students (entering freshmen class of 2009 and 2010) will be restricted by students who have been admitted to the institution. These cohorts will be further defined as: (a) students who have been admitted prior to financial aid offerings (in the case of the subject institution, this will involve students who have been admitted through the month of January of the given recruitment cycle); (b) students who were not recruited as athletes; and (c), students who have filled out the student profile questionnaire information from the SAT test (the common college entrance exam administered to Texas State college bound students). These restrictions will provide a cohort which has yet to be influenced by financial aid offerings within their decision making process, students who are not recruited differently due to athletic recruitment processes, and have consistent profile information based upon the questionnaire included in the SAT test. The later information from the SAT questionnaire includes such demographic information as the student's home address, their level of parental postsecondary education (if applicable), etc. Additionally, the SAT test and the students transcript data will provide information about the individual high schools and the size of the high schools that these students attended, as well as when they took the SAT test in their college attainment cycle.

These data sets will provide the information used as variables in this multivariate analysis models for determination in predictive modeling. The entire student profile build will include the independent variables for the purpose of the

study, with the dependent variable being defined as a discrete variable of whether or not a student enrolls in this particular Texas institution of higher education. The DesJardins method reviews the independent variables hypothesized affect on a student's probability of enrollment based upon each student's personal, background, and educational characteristics, and college preferences and intentions. Also included in this information will be the size of the student's high school and the enrollment to application "yield" that this institution has realized in previous years from the particular high schools in general. The selections of the variables used in this study are developed from surveying the relevant literature on student-choice (DesJardins, 2002).

Data Management

The sources of data which are used within the context of this study will exclude personally identifiable characteristics of the member of the entering cohorts. In addition, the design of this study does not necessitate a survey of the student's for the purposes of generating data. In analyzing and assessing the data from this study, student disclosure and student participation are restricted. The data analysis which will take place throughout the course of this study is done in aggregate; the design is such that the behaviors of the entering cohorts are grouped and therefore are not necessarily prescribed to specific student interaction. The purpose of the study seeks to review and assess the behavior of the entering cohorts of students and will not be used to discriminate or identify specific students throughout the course of the study.

CHAPTER IV

DATA ANALYSIS

In pursuit of reproducing the DesJardins model of predictive enrollment, the analysis of two entering classes of freshmen to the study institution were reviewed. In order to attempt to recreate a similar context study for evaluation, steps were taken to find similarities from the DesJardins model in the type of population sample, parallels between variables, and recreate similar methodologies for assessment of the data discovered. Though this study fails to reach a level of actual replication of this original study, the methodology for analysis within this study attempts to approximate the methodologies recommended by DesJardins (2002).

Two independent samples of entering freshmen classes of a public, four-year institution of higher education in the southwest were isolated and reviewed via analysis of predictive modeling for prospective student enrollment. The enrollment behavior was identified for the members of each of these entering classes as either "enrolled", or "otherwise". The hypothesis of the study attempts to predict this behavior by applying a logistic regression model; that is, does the predicted behavior match that of the observed behavior through the application of the logistic regression? The choice of logistic regression is an appropriate method designed to "model the relationship between one or more predictor variables and an outcome" (Field, 2005, p. 218); especially designed to facilitate the use of and predict outcome variables which are categorically dichotomous through the analysis of predictor variables that are

continuous or categorical. These specific limiting criteria render the application of linear regression/multiple regression modeling ineffective due to the lack of a linear relationship between the predictor(s) and the outcome variables.

In the DesJardin (2002) methodology, the researcher built a model for logistic regression predictive modeling by taking a specific year of data (i.e., single entering class) and randomly splitting the data set into two different groups. The first group was designed to be the "model development" group, while the second group was withheld to be a "hold out" or "validation model" to test the efficacy of the predictive model designed in the "developmental group". Similarly, this study takes the enrollment data from the entering freshmen class of 2009 and split this data into two, random groups of cases. The admitted freshmen class of 2009 was randomly coded into two groups, with an emphasis on equalizing the sample size of these two groups (n= 4,803, n= 4,805, respectively). This method independently aggregates a logistic regression model for both of these groups for comparison.

The samples of the entering classes of both 2009 and 2010 were analyzed with the inclusion of specific variables which are ubiquitous to the entering freshmen's characteristic data and available across multiple years of applications to the study institution. This data ultimately is derived from two sources, the ApplyTexas application (formerly the Texas Common Application, which includes demographic information) and the SAT college entrance exam Student Questionnaire, which is a voluntary survey component questionnaire included in the administration of this collegiate entrance exam. The following tables provide the definition of the variables

in the study, and the descriptive statistics within the entering class of 2009 sample data:

Table 1 Definition of the Variables

Variable	Description
Enrolled	1 if enrolled, 0 otherwise
African American	1 if African American, 0 otherwise
American Indian	1 if American Indian, 0 otherwise
Hispanic/Latino	1 if Hispanic/Latino, 0 otherwise
Other Ethnicity	1 if Other Ethnicity, 0 otherwise
White	1 if White, 0 otherwise (reference group)
Male	1 if Male, 0 if Female
Low Income	Family income less than $40,000
Middle Income	Family income between $40,000 and $80,000
High Income	Family income above $80,000 (reference group)
Admission Index Score	[(2*SAT Composite) + high school rank %
High School Size	Number of students in the high school attended
High School Yield	Historic Yield (enrolled/applied) from high school
TX Resident	1 if student is a Texas Resident, 0 otherwise
Parents Alumni	1 if either/both parent are study institution alumni, 0 otherwise
Public High School	1 if student attended a public high school, 0 otherwise
App before Oct.	1 if student applied before October, 0 otherwise
App in Oct.	1 if student applied in October, 0 otherwise (reference group)
App in Nov.	1 if student applied in November, 0 otherwise
App in Dec.	1 if student applied in December, 0 otherwise
App in Jan.	1 if student applied in January, 0 otherwise
App in Feb.	1 if student applied in February, 0 otherwise
App after Feb.	1 if student applied after February, 0 otherwise

Table 2 Descriptive Statistics of the Entering Class of 2009

Variable	N	%	Mean	S.E.	Min.	Max.
Enrolled	4,243	42.3				
African American	519	5.2				
American Indian	72	0.7				
Hispanic/Latino	1,673	16.7				
Other Ethnicity	0	0.0				
White	6,286	62.6				
Male	5,067	50.5				
Low Income	1.631	16.3				
Middle Income	2,338	23.3				
High Income	6,046	60.4				
Admission Index	9,608	--	2,329.48	2.370	1,325	3,290
High School Size	9,753	--	439.08	2.890	4	1,402
High School Yield	10034	--	41.14	0.224	0.0	100
TX Resident	9,479	94.5				
Parents Alumni	239	2.4				
Public High School	8,934	89.0				
App before Oct.	283	2.8				
App in Oct.	974	9.7				
App in Nov.	1,845	18.4				
App in Dec.	1,390	25.0				
App in Jan.	2.505	25.0				
App in Feb.	1,432	14.3				
App after Feb.	1,605	16.0				

Some of the variables have been established into “dummy variables” which control for redundancies of sampling; those variables have been labeled in “Definition of Variables” as the reference group.

The choice of variables for this model was considered for specific factors. First, the study is designed to approximate the same variables used in the DesJardins (2002) study. Some of these variables were either not existent in the SAT Student Questionnaire (the DesJardins model uses the questionnaire which accompanies the ACT college entrance exam), or they were not collected by the study institution. Second, the literature and research in the field of enrollment management provides correlation tendencies to the variables considered. Finally, the consistency of these

variables across large sample sizes provides for a robust data source for consideration of the model's application (N = 9,608).

The choices in application specified data reflect the enrollment cycle of the study institution. It is the institution's choice to move the enrollment application process earlier into the year to facilitate a better understanding of the predictability and planning for future enrollments. The application became available in the two enrollment cycle years (2009, 2010) at the beginning of September in the respective calendar years. The future application priority deadline will be established on March 1st of the given enrollment cycle; it is the study institution's desire to interpret these application trends in light of future enrollment predictions.

Data Analysis Procedure

The analysis of data used in this study began with the analysis of the entering freshmen class of 2009, using the aforementioned variables. In order to review the predictability of the logistic regression model, this class of 2009 was split into approximately two equal, random distributions of the samples within the 2009 class (missing cases withheld). Purposely, these two groups were differentiated to mitigate bias within the model testing/replication. There is a problem exposed to the model generated by the logistic regression if the sample is used for model development and model testing and tested for predictive accuracy; this produces an inherent bias towards a successful model (SAS Institute, Inc., 1995). The distinguishing of two groups of this original data set into a "developmental" group and a "validation" group

(referred to as Developmental Group and Validation group, respectively) provides some control over the use of the same cases for model building and model testing.

Data analysis of the sample (developmental group) was conducted using the "Binary Logistic" regression method within the SPSS (PASW) v.18 software suite. Additionally, the logistic regression method was also applied to the Validation Group in order to compare the analysis between groups. The Developmental Group, however, was used to determine the logistic regression equation which was incorporated into the Validation Group (holdout group) for comparison, and was incorporated into the Prediction Implementation Group (entering class of admitted freshmen in 2010) which occurred in the subsequent term. The variables (previously listed) where used within the model as predictors of the outcome variable (Enrolled student) for the determination of the effectiveness of the model and to maintain continuity.

The method of logistic regression chosen for the analysis in this model was the "forced entry method", or "enter" method as it is referred to in the software suite. Choosing the forced entry method was based upon a twofold determination. First, there are several numbers of dummy variables in place among the variables used in this study. Since a number of the variables are categorical, "the obvious problem with wanting to use categorical variables as predictors is that you'll have more than two categories" (Field, 2005, p. 208). Field (2005) further notes that all related dummy variable coding must be entered in the same block for analysis (there is discontinuity of adding dummy variables in a stepwise regression analysis, since the dummy

variables are all related within a specific predictor as a whole). To this end, the categorical dummy variable inclusion was held in the forced entry method to ensure that the dummy variables within a predictor's case would be held constant to the overall predictor's effect. Second, Field (2005) also notes that when testing a model theory, the forced entry method is best used in this scenario (whereas stepwise regression analysis is better suited to theory exploration). This study sought to incorporate the forced entry method with all of the variables in the first block of entry to maintain the integrity of all predictors being equal in the analysis.

The analysis design involved specifying the outcome dichotomous categorical variable to be the actual enrolled student of the admitted student class. The remaining variables were established to be predictor variables within the model. Specific, optional, statistical methods were chosen through the process in order to analyze the data. These include, among other, the Hosmer-Lemeshow goodness-of-fit test, casewise listing of the residuals, and the recording of the Confidence Intervals (CI) of the exponetiated *b* values for each predictor variables. These statistics will be discussed below. Additionally, the cutoff classification value was adjusted from the default (0.5) to a stricter threshold of (0.4) in order to decrease the likelihood of incorrectly classifying actual enrolled students, as discussed and recommended under the DesJardins method (2002). Though this increases the likelihood of incorrectly classifying non-enrolled students, in enrollment management the desire is to err on the side of the enrolled student since this is the favorable outcome group of the most interest.

Within the logistic regression analysis, the first case processing summary classification table of the model only includes the constant in the regression equation (i.e., all of the predictor variables are omitted). This opening analysis sequence compares the model data versus the observed data in the simplest of terms and determined that the predictability of the model with only the constant inclusion in the equation to provide only 43% of the model determination correct in comparison of the observed data. Therefore, withholding all other information (predictor variables), the likelihood of correctly predicting enrollment of an incoming class through the use of this sample data is found to be slightly under the probability of the actual event occurring due to chance (50%). This opening sequence also establishes the -2 × *log-likelihood* value of the constant (-2LL = 6,562.628) within the equation which will test the inclusion of the predictor variables in the subsequent iterations. The *log likelihood* value is the logarithmic expression of *likelihood* in a regression equation, synonymous to the *residual sum of squares*, or level of unexplained variance – which is a function of the sum of difference between predicted and observed values (Field, 2005). SPSS produces a *log likelihood value* multiplied by two in order to maintain a chi-square distribution on the score for ease of comparison and study.

Developmental Group

Evaluation of the logistic regression statistic development provided more information on the choice of variables and the predictive ability for the prospective model. As seen above, the baseline model for this study predicted approximately 43% of the outcome and had a very high value for the -2 × *log likelihood* measure. The

opening analysis of the logistic regression methodology also provided initial interpretation on the model variables withheld from the equation during baseline generation. These variables generated a residual chi-square statistic which was significant at $p < .001$. This value indicated that the coefficients for the variables withheld from the baseline model are significantly different from zero; one or more of these variables (with respective coefficients) will significantly affect the predictive power of the logistic regression model.

After the establishment of the baseline information, the logistic regression implemented again with the inclusion of the variables for the model development. Notably, the first indication of successful interpretation of the included predictors in the model is found in the -2 × *log likelihood* (-2LL) value. The opening, constant held model without the inclusion of the variables in the equation (-2LL = 6,562.628) was decreased with the inclusion of the variables to the equation (-2LL = 5,308.903). The model chi-square statistic on the *Omnibus Tests of Model Coefficients* provides confirmation that the coefficients derived from the variables within the model are significantly different from the baseline model and provides a better model of prediction with the inclusion of these variables in the model development process (19 DF, $p < .001$). Field (2005) indicates that the model chi-square score/significance is an analogue to the *F*-test for the linear regression's sum of squares; as such, this significant model chi-square provides support to the improvement in the predictability of the model with the inclusion of the variables compared to the inaccuracies inherent within the model itself.

Additionally, the logistic regression of the developmental model provided information regarding the nature of the variables introduced into the model. Estimates for the coefficients of each variable were created within the model which would be used in the prediction equation for both model validation and implementation for future iterations. The *b*-value for each variable indicates the coefficient for each variable. This coefficient value of a given predictor provides an indication of change within the outcome resulting from a unit change within each individual predictor variable. In logistic regression, this represents a change in the logit of the outcome variable per one-unit change in the predictor variable. The logit represents change of the natural logarithmic odds ratio of the given *Y* (probability) of outcome as influenced by each predictor variables value (Field, 2005). These values are described in the *Variables in the Equation* SPSS output found in Table 3.

With the estimated coefficients generated with this model, the SPSS output also provided scoring of the Wald statistic. The Wald statistic represents a chi-square distribution and determines whether or not a given predictor's *b*-value coefficient for a given predictor variable is significantly different from zero. If this statistic is significant, then it is safe to assume that the predictor is providing contribution to the prediction of the outcome. The Wald statistic in logistic regression is analogous to the *t*-statistic used in linear regression; as such the coefficient is evaluated through the standard error of the coefficient:

$$\text{Wald} = \frac{b}{\text{SE}\, b}$$

The Wald statistic is a ratio determined by the coefficient divided by the standard error of the coefficient being evaluated. A significant Wald statistic was found for the variables: Application before October vs. Application after February, Admissions Index Score, High School Yield, and Parent(s) Alumni status (all less than the .05 level).

The tertiary component of the *Variables in the Equation* table from the SPSS output is found within the value of the exponetiated *b* (Exp(B) for each variable introduced into the model. This value is an indicator of the change in odds with the inclusion of the given predictor variable unit change. This tool, especially useful in the evaluation of the change in odds when using categorical variables, represents the change in the probability of the event occurring (enrolling) versus the probability of the event not occurring (not enrolling, or otherwise) given the included predictor:

$$\text{odds} = \frac{P(\text{event})}{P(\text{no event})}$$

A positive Exp(B) score (greater than 1.0) indicates that an increase in the predictor variable's unit change will be an odds increase; a score less than 1.0 indicates the inverse. The Exp(B*)* value's confidence intervals are included in the data reported. The confidence intervals encompass 95% of the possible values for the variable within the upper and lower limits listed. Variables which have upper and lower limits contained as either greater than 1.0, or less than 1.0 values (without crossing 1.0) provide corroboration of the effect of the odds ratio changing statistically in the Exp(B) value. Confidence intervals which span across the 1.0 interval indicate that the variable produces an increasing or decreasing effect on the odds ratio and cannot

be effectively be contained in the model since their impact is unknown. As seen in the Wald statistic earlier, the variables for Application Date prior to October versus Application Date after February, Admissions Index Score, High School Yield, and Parent(s) Alumni affiliation are the only Exp(*b*) values which have confidence intervals contained within the interval limits of greater than or less than 1.0, without spanning the interval of 1.0 (with the lower limit < 1.0 and the upper limit > 1.0, respectively). Only the High School Yield variable demonstrated a positive change per unit increase of the predictor variable when assessing change within the odds of enrollment.

Finally, interpretation of the effect of including the variables in the model development demonstrated successful influence upon the probability of predicting enrollment. The original probability of enrollment derived from the constant model (withholding the variables in the equation) determined that the model could only predict the probability of enrollment at 43%. With the inclusion of the variables in the model (in the equation), the probability of predicting enrollment increased to 68.5%. This analysis was conducted with a cutoff value of .400, limiting the predictive nature of the analysis in favor of underestimating the prediction of enrollment, since this is the most crucial value in consideration. The error likelihood more likely favors a Type I error for this study; where the error would be skewed more towards not including an enrolled student instead of incorrectly including a non-enrolled student in the probability decision.

Table 3 Variables in the Equation

Step 1	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for Exp(B)	
							Lower	Upper
Ethnicity1	.019	.160	.014	1	.907	1.019	.745	1.393
Ethnicity2	.668	.400	2.788	1	.095	1.950	.890	4.370
Ethnicity3	.132	.102	1.684	1	.194	1.141	.935	1.392
Ethnicity4	.102	.098	1.092	1	.296	1.108	.914	1.384
FamilyIncome1	.156	.086	3.289	1	.070	1.169	.987	1.384
FamilyIncome2	-.063	.104	.370	1	.543	.939	.765	1.151
Application1	-.188	.221	.727	1	.394	.829	.538	1.277
Application2	.157	.297	.571	1	.450	1.169	.779	1.755
Application3	.215	.213	1.012	1	.315	1.239	.816	1.883
Application4	-.009	.203	.002	1	.966	.991	.666	1.476
Application5	.105	.213	.243	1	.622	1.110	.732	1.685
Application6	-.574	.211	7.412	1	.006	.563	.373	.851
Male	-.007	.068	.011	1	.916	.993	.870	1.134
AdmissionsIndexScore	-.001	.000	21.529	1	.000	.999	.999	1.000
HighSchoolSize	.000	.000	1.173	1	.279	1.000	1.000	1.000
HighSchoolYield	5.338	.203	694.706	1	.000	208.010	139.864	309.357
TexasResident	-.342	.185	3.401	1	.065	.711	.494	1.022
ParentsAlumni	-.811	.222	13.291	1	.000	.445	.288	.687
AttendPublicHS	.003	.135	.001	1	.982	1.003	.770	1.306
Constant	-1.036	1.132	.837	1	.360	.355		

Model Assessment and Goodness-of-Fit

The Hosmer and Lemeshow Test for goodness-of-fit provides information not only about the effectiveness of the model, but also the correlation of the coefficients within the equation. Through use of this statistical method, a model R^2 correlation coefficient (the coefficient of determination) was determined for the developmental group to account for total variance in student enrollment by the introduced predictor variables. The Hosmer and Lemeshow (Hosmer & Lemeshow, 2000) method for calculating R^2 involves the model chi-square distribution divided by the original -2LL of the model; invoking the results from the developmental model provides a model chi-square value (1,253.724) divided by the original -2LL value (6562.628) for a Hosmer and Lemeshow R_L^2 value of .191, or approximately 19% of the variance in the enrollment probability. Additionally, the SPSS logistic regression methodology produces two R^2 estimates; the Cox & Snell's R_{CS}^2 (.230 in the developmental model) and the Nagelkerke's R_N^2 (.308) respectively. When evaluating all three of the R^2 estimates, only a relatively small portion of the variance in enrollment probability can be attributed to the incorporation of the variables within the developmental model.

The Hosmer and Lemeshow Test more importantly provides analysis of the goodness-of-fit for the developmental model when describing the inclusion of the variables. This statistic "tests the null hypothesis that the model fits the data (thus, a significant HL test provides evidence that the model does not fit the data)" (DesJardins, 2002, p. 541). In the developmental model, the Hosmer and Lemeshow

Test score is 25.742 with 8 degrees of freedom (p = .001). Attenuating the significance of the Hosmer and Lemeshow Test to the developmental group's model assessment, the criteria of non-significance is not met with the test and the model is found to not accurately fit the data.

In addition, the Hosmer and Lemeshow Test allocates the outcome predictors into deciles for evaluation. These deciles provide a comparison of the observed outcome versus the predicted outcome by separating the sample out into to ten equitable groups. The purpose of this design provides not only a comparison across large samples, but also provides a sequential segmentation of probability as a ratio of tenth units. This step provides the research with a *de facto* score consistent to the ratio of probability on a percentile scale. The contingency table provides some analysis of the model where observed and expected data are close; this model tended to predictor better expected results in the upper deciles. The SPSS Hosmer and Lemeshow contingency table is provided below:

Table 4 Contingency Table for Hosmer and Lemeshow Test – Developmental Group

Step 1	Student Enrolled in Institution = 0		Student Enrolled in Institution = 1		Total
	Observed	Expected	Observed	Expected	
1	463	443.856	17	36.144	480
2	407	398.620	73	81.380	480
3	341	364.060	139	115.940	480
4	316	332.509	164	147.491	480
5	290	302.886	190	177.114	480
6	266	271.972	214	208.028	480
7	253	237.627	227	242.373	480
8	199	196.785	281	283.215	480
9	144	135.863	336	344.137	480
10	61	55.822	422	427.178	483

Hosmer-Lemeshow goodness-of-fit statistic = 25.742 with 8 DF (p = .001)

Validation Model

The logistic regression of the developmental model provided the elements to construct a model regression equation for testing and replication. Building upon the derived equation, the second group of the entering admitted freshmen class of 2009 provided a sample for use as a Validation Group for this study. By taking the validation sample and overlaying the Developmental Group's logistic regression equation, probabilities for enrollment are developed for analysis and cross validation.

For the purpose of comparing the Validation Group to the Developmental Group, the basic logistic regression equation is used. Substituting the estimated coefficients from the Developmental Group gave the Validation Group a computed value, the Modified Score, for use in the logistic regression analysis. The equation is defined as:

$$log\ \frac{P}{1-P_i} = b_0 + b_1X_1 + b_2X_2 \ \ldots\ b_iX_i$$

where the logistic regression is the summation of the constant (b_0) and the coefficients of each of the predictor variables (b_iX_i) across all of the predictor variables. Using the estimated coefficients of the Development Group and substituting them for the coefficients of the Validation Group provided scores for each predictor variable used in the logistic regression and testing of the model. In this logistic regression model, each case within the Validation Group had a Modified Score variable computed.

The opening analysis of the Validation Group, the constant group analysis, demonstrated the same classification probability for enrollment prediction of 43% (as expected). The -2 × *log likelihood* of the constant group (Validation Group -2LL = 6,566.007) was similar to the Developmental Group (Developmental Group -2LL = 6,562.628) and the introduction of the predictor variable of the Validation Group in the equation reduced the -2 × *log likelihood* with inclusion of the variable (-2LL = 5,634.792, with Modified Score variable in the equation), increasing the predictive nature of the model with inclusion by a probability of predicting enrollment increasing to 65.1%. The model chi-square score was significant ($p < .001$), indicating that the variable (Modified Score) was a significant indicator for increasing the predictive nature of the logistic regression.

Inclusion of the Modified Score predictor variable in the equation provided valid additional information for the equation. Specifically, the Wald Statistic was found to be significant ($p < .001$), with an exponetiated *b* **(**Exp(B) value indicating a positive relationship towards prediction of the coefficient per unit-change of the

predictor variable. Additionally, the Confidence Interval for Exp(B) were inclusive values greater than 1.0, verifying the positive influence of the coefficient in the equation, and only a 5% chance that the confidence interval does not include the Exp(B) value range. Because the Exp(B) value demonstrates a positive relationship to the unit-change in the predictor variable, and because the value is constrained within the Confidence Intervals (both greater than 1), then it is foreseeable to assume the introduction of the Modified Score variable to the logistic regression is generalizable to the greater population of admitted students. Table 5 provides a summary of the variable introduced into the equation.

Table 5 Modified Score Variable in the Equation for Validation Group

							95% C.I. for Exp(B)	
Step 1	**B**	**S.E.**	**Wald**	**df**	**Sig.**	**Exp(B)**	**Lower**	**Upper**
ModScore	.859	.033	680.624	1	.000	2.361	2.213	2.518
Constant	-.262	.032	66.176	1	.000	.770		

Validation Group Assessment and Goodness-of-Fit

The Validation Group provides assessment measures which are examined in the same manner as the Developmental Group. The Hosmer and Lemeshow method for calculating the R^2 value for the sample involves the model chi-square score of the Validation Group (931.215) divided by the original -2 × *log likelihood* value of the model held to only the constant in the equation (-2LL = 6,566.007); this calculation

produces the model R_L^2 of .142, or approximately 14% of the variance within the model. Similarly, the analogous R^2 methods produced in the SPSS analysis, the Cox & Snell R_{CS}^2 and the NagelkerkeR_N^2, produce R^2 values of .176 and .236 respectively. These later two tests tend to inflate above the Hosmer and Lemeshow R^2 method, but the overall analysis of R^2 is held around 18% variance within the model with the included variable in the equation.

The Hosmer and Lemeshow methodology additionally includes the goodness-of-fit test within the logistic regression model. In the Validation Group, the Hosmer and Lemeshow goodness-of-fit statistic provided a chi-square value of 35.898 with 8 degrees of freedom ($p < .001$). Again in this model, the significance of the Hosmer and Lemeshow goodness-of-fit is in itself significant, which detracts from the actual fit of the model to predictability – the same problem found in the Developmental Group model's goodness-of-fit evaluation. Likewise, the Hosmer and Lemeshow method provides for a segmentation of the predictability of the sample's cases by dividing them into deciles and comparing the observed versus predicted values based upon the logistic regression (Table 6).

Table 6 Contingency Table for Hosmer and Lemeshow Test - Validation Group

Step 1	Student Enrolled in Institution = 0		Student Enrolled in Institution = 1		Total
	Observed	Expected	Observed	Expected	
1	453	426.887	28	54.113	481
2	350	379.775	131	101.205	481
3	338	349.595	143	131.404	481
4	317	322.609	164	158.391	481
5	284	298.901	197	182.099	481
6	280	272.441	201	208.559	481
7	241	243.338	240	237.612	481
8	221	207.915	260	273.085	481
9	184	161.750	297	319.250	481
10	72	76.738	404	399.262	476

Hosmer and Lemeshow goodness-of-fit statistic = 35.898 with 8 df ($p < .001$)

Model Testing

After examining the Developmental Group through a logistic regression, an equation for the predicted probability for enrollment was developed. Taking this equation and applying it to the Validation Group provided a logistic regression which was found to both significantly increase the predictive nature of the logistic regression in the Validation Group as well as explain some of the variance found in the model logistic regression scenario. Building upon this premise that there was a benefit to applying the Developmental Group's equation to the Validation Group, a means of analyzing the effectiveness of comparing the predictive nature of these two groups was attempted.

In order to assess the effectiveness of applying the Developmental Group's logistic regression equation onto the Validation Group's sample, correlational analysis

of the logistic regression equation and the predicted outcome variable was conducted. In the case of evaluating these two groups, the most effective method of correlational study derives from analysis of the point-biserial correlation coefficients of the two sample groups. The point-biserial correlation coefficient (r_{pb}) method was chosen since the outcome variable in both instances is dichotomous and discrete (i.e., the outcome variable of enrolled is either "enrolled" or "otherwise", there are no other classifications of the variable) (Field, 2005).

In the analysis of the point-biserial correlation coefficient, the test was conducted using the Pearson's Correlation with a one-tail test for significance. The test produces a *r* value for both the Developmental Group and the Validation Group for comparison. These two correlations are listed in the tables below:

Table 7 Point-Biserial Correlation - Developmental Group

		Student Enrolled in the Institution	ModScore
Student Enrolled in the Institution	Pearson Correlation	1	.419
	Sig. (1-tailed)	-	.000
	N	4803	4803
ModScore	Pearson Correlation	.419	1
	Sig. (1-tailed)	.000	-
	N	4803	4803

** Correlation is significant at the .01 level (1-tailed).

Table 8 Point-Biserial Correlation - Validation Group

		Student Enrolled in the Institution	ModScore
Student Enrolled in the Institution	Pearson Correlation	1	.419
	Sig. (1-tailed)	-	.000
	N	4805	4805
ModScore	Pearson Correlation	.419	1
	Sig. (1-tailed)	.000	-
	N	4805	4805

** Correlation is significant at the .01 level (1-tailed).

The point-biserial correlation coefficient (r_{pb}) can be squared in order to produce a R^2 value. In the case of the Developmental Group and the Validation Group, this value is essentially the same (R^2 = .176). Cross-validation of regression analysis provides a means of comparing the two regression models; in this case, the original Developmental Group (the equation derivative group) and the Validation Group (the model test group) can be compared through the comparison of the two respective R^2 values. Since the R^2 values are the same, the cross-validation confirms the effectiveness of the logistic regression equation used within the Validation Group.

ROC Curve and Brier Score

Following the DesJardins methodology, two more *post hoc* evaluation methods were conducted upon the Validation Group data in an effort to ensure the effectiveness of the predictive model generated through the development of a logistic regression

equation for predicting enrollment. The first method consists of constructing a receiver operating characteristics (ROC) curve as a graphical analysis. The second method, the Brier Score, provides an additional measure of how a model predicts the accuracy of events. These two methods were originally conducted in the DesJardins study and are replicated here.

The receiver operating characteristics (ROC) curve is "a graphical way to test the predictive accuracy of a logistic regression" (DesJardins, 2002, p. 541). This curve plots the sensitivity versus one minus the specificity of the estimated model. In the case of this study, the sensitivity (selectivity) represents the percent of correctly predicted enrollees (versus non-enrollees). The evaluation of the test is represented by a diagonal, 45-degree line upon a graph, and produces a "c" statistic as a value of evaluation of the model. The "c" statistic represents a percentage score of the predictive nature of the logistic regression, where a score of 0.5 would be a result no better than a chance incident, and a value of 1.0 would indicate perfect prediction of the observation. Graphically, a ROC curve would demonstrate a higher degree of prediction as the curve tends to bow towards the upper left-hand corner of the graph. The "c" statistic is a representation of the area within the curve extrapolated out from the 45-degree test line. The Validation Group ROC curve is the following figure:

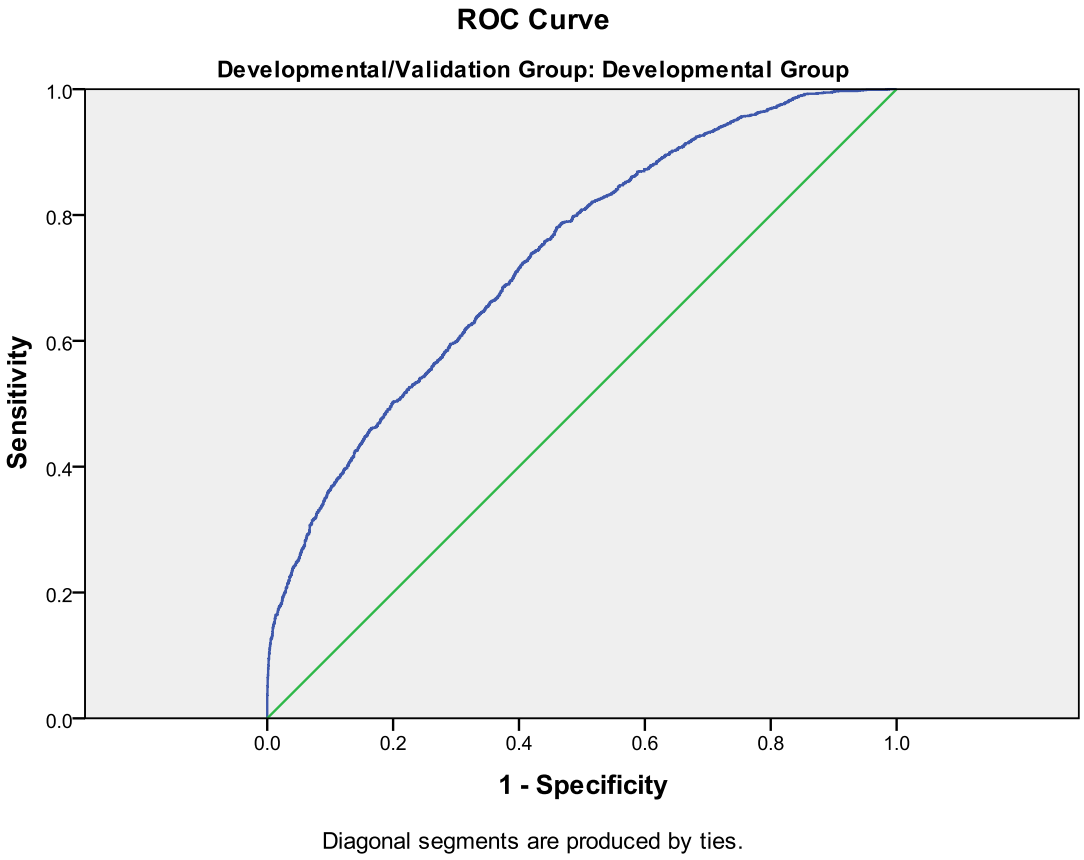

Figure 2 Validation Group ROC Curve

The "c" statistic for the Validation Group ROC curve is .733 (where the null hypothesis would equal "c" = 0.5). The "c" statistic value of .733 is considered a good value for predictive accuracy.

The Brier score (Brier, 1950) derives from evaluating the predictive accuracy of weather forecasting. In logistic regression, the Brier score is a mathematical *post hoc* evaluation based upon the predicted outcomes and the observed outcomes. The equation for this method demonstrates the average of these total predictions and observations:

$$\text{avg}\ [(p_i - y_i)^2]$$

where (p_i) is the predicted value and (y_i) is the observed value. Taking the probabilities of each case within the sample and comparing them to the observed outcome, a value of variance for each case within the sample can be computed according the Brier score method. These values can then be averaged in order to obtain a Brier score for the Validation Group. The Brier score "is a strictly proper scoring rule, which means that it is minimized for predicted probabilities that are equal to the true probabilities" (SAS Institute, Inc., 1995, p. 35). In evaluating the Brier score, the index of the score is a value between 0 and 1, with smaller scores indicating better predictive accuracy. The Brier score for the Validation Group was found to be .69 for the model.

Validation Model Summary

Through the use of logistic regression analysis, a model was developed (Developmental Group) which provided a logistic regression equation for use in predicting enrollment of an entering group of freshmen at the study institution. This model was tested against a holdout sample (Validation Group) in order to evaluate the predictive accuracy and applicability of the logistic regression equation. The Validation Group provided evidence of the effectiveness of the model, though the overall predictive nature of the model was marginally effective and generalizable. A summary of the logistic regression of the Validation Group is as follows:

Table 9 Validation Group Summary

		95% CI for Exp(B)		
	***B* (SE)**	**Lower**	**Exp(B)**	**Upper**
Included				
Constant	-.262			
	(.032)*			
Modified Score	.859	2.213	2.361	2.518
	(.033)*			

Note R^2 = .14 (Hosmer & Lemeshow), .18 (Cox & Snell), .24 (Nagelkerke).
Model x^2 = 931.22, p < .001, * p < .001
Hosmer & Lemshow goodness-of-fit statistic = 35.898 with 8 DF (p < .001)

Cutoff (prior probability)	40.0
Correct Classification rate	65.1
Brier Score	0.69
"c" statistic	0.73

Predictive Model Implementation

Finally, after developing a model equation and testing the validity of the model, the next step in the process of applying logistic regression analysis for the purposes of predicting enrollment for an incoming class involved using the model in a future enrollment scenario. At the time of this study, the entering class of 2010 had just matriculated to the study institution. This timing allowed for the actual analysis of the predictive nature of the developed model logistic regression to truly be analyzed. Building upon the developed logistic regression equation, a sample of the incoming admitted freshmen class of 2010 (N = 6,035) from the study institution was subjected

to the logistic regression analysis with the Developmental Group's logistic regression equation. For the purposes of comparison, the same parameters for defining the sample were held consistent in the 2010 class sample data; that is, the class included only cases where the student was admitted to the subject institution, completed the SAT college entrance exam optional Student Questionnaire information, and held values within the predictor variables identified in the study in the previous administrations of the logistic regression analysis. In the event that a case did not have the specific predictor variables/the specific predictor variables were missing, than the individual case was excluded from the sample. All efforts were made to maintain the closest continuity of sample information for the purposes of implementing this predictive enrollment model.

The entering class of the fall of 2010 (the subsequent enrolling class after the Developmental/Validation Group) was subjected to the SPSS logistic regression analysis with the introduction of the Modified Score variable as a predictor variable; enrolled or "otherwise" being the outcome variable. The Modified Score variable introduced was similar to the Modified Score variable produced in the Validation Group; that is, the logistic regression equation's coefficient values were the original values discovered in the Developmental Group's primary logistic regression analysis. Substituting the Developmental Group's coefficient values with the class of 2010's own unique predictor variable values per case applied the predictive enrollment model to the new data set. Once the Modified Score predictor variable value was computed,

then the logistic regression analysis was conducted using Enrollment as the outcome variable and Modified Score as the predictor variable.

The predictive enrollment model on the class of 2010's logistic regression analysis produced similar results to those of the development and hold out samples created earlier. The first iteration of the logistic regression analysis (using only the constant of enrolled or otherwise) demonstrates a predictive accuracy of approximately 41.1%. This initial analysis demonstrated a very high -2 × *log likelihood* value (8,172.330) with the constant, as expected, being significant ($p <$ .001). The addition of the Modified Score predictor variable being added to the logistic regression demonstrated both an increase in predictability of the model (probability of correctly predicting enrollment rose to 68.4% and a decrease in the -2 × *log likelihood* value (-2LL = 6,788.787, indicative of an increase in the accuracy of predicting the outcome of "enrolled"). The model chi-square distribution was significant ($p < .001$), indicative of the confidence that the introduction of the Modified Score predictor was conducive to the predictive increase seen in the model. Though these *log likelihood* values are still very large, the impact of introducing the predictor to the logistic regression does increase predictability.

Analysis of the introduction of the predictor variable into the equation provided additional confirmation to the significance of the Modified Score predictor to the equation. The Modified Score predictor had a significant Wald statistic ($p < .001$) for the predictor's coefficient, and the exponetiated *b* (Exp(B) = 2.444) value was held within a very narrow Confidence Interval which was constrained on both the upper

and lower limits to values greater than 1.0; which indicated a positive relationship for the Modified Score predictor's coefficient. A per-unit change within the predictor variable was augmented by the positive lift from the predictor's coefficient, and the Confidence Interval values indicated that the 95% of the coefficient's variance would indeed compliment this positive relationship. Since the Confidence Interval (both upper and lower limits) were greater than 1.0, the generalizability of the Modified Score predictor's predictive nature can be concluded to be an enhancing contribution to the increased predictive nature of the logistic regression analysis.

Prediction Model Implementation Assessment and Goodness-of-fit

As seen in the earlier development and validation testing of the prediction model, the Prediction Model Implementation on the entering class of 2010 was subjected to the same battery of assessment and goodness-of-fit methodologies. Assessment of the model's predictor was conducted through the traditional correlation analysis methods. The model's analogous coefficient of determination (R^2 value) was computed through the Hosmer and Lemeshow method (R_L^2 = .169), or approximately 17% of the variance attributed to the introduction of the predictor variable. Similarly, the SPSS software computed the additional R^2 values; the Cox & Snell value (R_{CS}^2 = .205) and the Nagelkerke value (R_N^2 = .276) for the coefficient of determination. Though small, the model's R^2 value(s) indicate that there is a correlation of the variance in the predictive nature of the model due to the introduction of the Modified Score predictor variable.

The Hosmer and Lemeshow goodness-of-fit statistic also portrayed a similar trend in assessment as seen in the model development. The Hosmer and Lemeshow goodness-of-fit statistic was 37.285 with 8 degrees of freedom ($p < .001$). Like seen in the development of the model, the significance of the Hosmer and Lemeshow goodness-of-fit test detracts from the model (where a non-significant Hosmer and Lemeshow goodness-of-fit value would indicate collusion between the observed and the predicted models). The contingency table for the Hosmer and Lemeshow goodness-of-fit method demonstrates the predictive nature of the model (Table 10). Again, this test provides a breakdown of the observed versus predicted outcome of the model by segmenting the sample into deciles.

Table 10 Contingency Table for Prediction Model

Step 1	Student Enrolled in Institution = 0		Student Enrolled in Institution = 1		Total
	Observed	Expected	Observed	Expected	
1	574	546.815	31	58.185	605
2	479	500.934	125	103.066	604
3	440	462.584	165	174.731	605
4	429	429.584	175	174.731	604
5	383	396.989	221	207.208	604
6	354	359.989	250	244.011	604
7	316	315.894	288	288.106	604
8	286	265.130	318	338.870	604
9	224	195.265	380	408.735	604
10	72	84.327	525	512.673	597

Hosmer & Lemshow goodness-of-fit statistic = 37.285 with 8 DF ($p < .001$)

The contingency table shows that there is some very strong predictive convergence, particularly in the seventh decile. Additionally, by controlling the cutoff value (.40), the tendency of the predictive model was to underestimate the predicted enrollment

versus the observed enrollment. Though there is question about the goodness-of-fit of the model, the demonstrated decrease in the *log likelihood* value, the Wald Statistic significance, and the R^2 value, which all trend toward an increase in the positive predictive nature of the Prediction Model. A summary of the Prediction Model logistic regression is as follows:

Table 11 Prediction Model Summary

		95% CI for Exp(B)		
	***B* (SE)**	**Lower**	**Exp(B)**	**Upper**
Included				
Constant	.809 (.047)*			
Modified Score	.894 (.029)*	2.311	2.444	2.585

Note R^2 = .17 (Hosmer & Lemeshow), .21 (Cox & Snell), .28 (Nagelkerke).
Model x^2 = 1,383.54, $p < .001$, * $p < .001$
Hosmer & Lemshow goodness-of-fit statistic = 37.285 with 8 DF ($p < .001$)
Cutoff (prior probability) 40.0
Correct Classification rate 68.4

CHAPTER V

CONCLUSIONS AND RECOMMENDATIONS

Predictive modeling methodologies provide a significant tool for interpretation and planning of enrollment cycles when applied to the study of student enrollment. Like all types of regression analysis, building upon observations of behavior, while controlling for the inclusion of various predictor variables, provides trending data which can be used to predict the probable outcomes of future events. In enrollment management, regression analysis provides for a method of evaluating the probability of trends of future enrollment behavior in incoming admitted students to a respective institution. This methodology elevates the predictive level beyond mere chance (50% enroll, 50% non-enrolled) and should be a tool for the any enrollment management professional to use in the development of recruitment practices customized to various groups of admitted students, to the estimates of potential class enrollment for institutions budgetary planning, and for resource allocation.

The purpose of this study involved the approximate replication of a similarly designed study created by Stephen L. DesJardins (2002) investigating the predictive potential of various data elements which are provided to an institution of higher education by prospective students through their application and submission of test score questionnaires in the application process. Taking the elements of variable information which could be expressed across the sample (i.e., continuity of variables across cases), DesJardins demonstrated a model for prediction of enrollment through

logistic regression analysis. This intriguing method provided the foundation for this study.

Model Adaptation

In design, the DesJardins method provided a template for replication which showed only some variances for this study with the available data of predictor variables. The study institution favored the SAT collegiate entrance exam over the ACT collegiate entrance exam, so some of the variables from the DesJardins method were not available. Wherever possible, the model developed by DesJardins was approximated with the data available in the study institutions population; the variables involved within the study were consistent the variables reviewed in the literature and research within the field of enrollment management. Based upon these elements and continuity, the sample data for the enrollment trends of the study institution were incorporated in the design and evaluation suggested by the DesJardins method.

Research Questions

This study revolved around the evaluation of predicting enrolling student classes based upon the logistic regression method. Specifically, the questions posed earlier in this discussion of the study centered on three factors. First, does the incorporation of predictor variables consistent across each case in a sample of admitted freshmen provide any indication to the probability of actual matriculation? Second, is the methodology used to generate a predictive model statistically sound and affective? Third, can the model derived from analysis in this study be used upon

subsequent enrolling classes to enhance the probability of predicting that a specific group of admitted students within the sample would enroll in the study institution? These questions were satisfied, in part, through the examination of the samples and the development of a logistic regression model.

The first question within this study examined the effect of the inclusion of various predictor variables consistent among cases in a sample population. The study evaluated the logistic regression model through three separate iterations, with three separate data samples (the model development sample, the model validation sample, and the implementation sample of a subsequent enrollment period). Through these three samples, the same prospective predictor variables were consistent across all of the groups. Using these predictor variables ensured the continuity of the logistic regression equation which was used in validation and implementation. The results of this equation's interpretation of probability of predicting enrollment were effective in increasing the potential probability of accurately predicting enrollment. In the case of the Developmental Group, the increase in probability of accurately predicting the enrollment of the student cases was increased from 43% to 68.5%. The logistic regression on the Developmental Group decreased the -2 × *log likelihood* values with the inclusion of the predictor variables, though these values were very high. The significance of the model chi-square test, the Wald statistic, and the interpreted R^2 values brought creditability to the use of the predictor variables in increasing the predictive nature of the model. When reviewing the model's predictors separately, there were few distinguishing variables which stood apart from the model with

positive (or negative) coefficient values to report as singular attributing predictors towards the increase in probability of prediction. Rather, the sum total impact upon the variance of the model with all included predictor variables provided indication that at least 19% of the variance in the outcome could be contributed to the total predictor variables interaction in the outcome. While the effect was measurable in the increased probability of prediction, there was still an unattributed amount of variance unaccounted for in the model. Coincidently, the model's goodness-of-fit test (the Hosmer and Lemeshow goodness-of-fit statistic) reduced the likelihood of the chosen variables serving as "best" predictor variables for consideration (demonstrated by the smaller threshold of the coefficient of determination). Nonetheless, the model did statistically increase the predictive nature of the logistic regression, providing valuable information for enrollment management professionals to use in the marketing and recruiting of these prospective admitted students.

As seen in the Developmental Group, the Validation Group and the Prediction Implementation Group both responded similarly in significance and predictability. The Validation Group increased the probability of accurately predicting enrollment from 43% to 65.1%; the Prediction Implementation Group increased its probability from 41.1% to 68.4% as well through the implementation of the model. These later two groups were conducted through the logistic regression using the model developed by the logistic regression in the Developmental Group. This model provided additional predictive power to the logistic regression analysis for each of these two later groups.

Just as observed in the Developmental Group, these later two groups demonstrated significance in their statistical analysis. Both of these groups saw a reduction in their -2 × *log likelihood* values, and both tested significantly in their chi-square analysis. For the samples used in the Validation Group and the Prediction Implementation Group, the complete listing of all predictor variables were rescored and computed into a new variable, the Modified Score variable. This variable represented the logistic regression equation, using the unique values of the predictor variables of each distinct sample, as the logistic regression predictor versus outcome variables. What is distinct about this analysis is that though there is a reduction in the total number of variables individually applied to the logistic regression analysis, the outcomes were similarly distributed across both of these later two models. Both the Validation and Prediction Implementation Groups had significant Wald statistic values (in this case, the Modified Score predictor variable) indicating that this coefficient was significantly different than zero. Both groups demonstrated exponetiated *b* values which were positive and maintained within 95% Confidence Intervals; these values indicated that the predictor Modified Score (in this case, the model logistic regression equation generated from the Developmental Group) increased the per-unit change in each predictor and, in turn, enhanced the probability of accurate prediction of enrollment. The coefficient of determinism (R^2) for each of these groups indicate that the model did indeed account for some of the variance in the increased probability of prediction; though, roughly at the same value as that seen in the Developmental Group (i.e., 14% to 16% respectively). Unfortunately, the goodness-of-fit demonstrated by

all of the regression analyses detracted from the statistical validation of the logistic regressions, leaving question as to whether or not the predicted model differs significantly from the observed data outcomes.

Building upon the significant findings from the logistic regression model development, the entering class of 2010 sample was subjected to the logistic regression equation generated in the model development phase of this study. The equation performed similarly on the Prediction Implementation Group, exhibiting the ratios in improved probability of predicting enrollment and a reduction in the *log* -2 × *likelihood values.* The Prediction Implementation Group did benefit from the logistic regression analysis; in terms of professional implementation, the model provided a better understanding and prediction of enrolling students than would be found in historical analysis. The importance of this implementation could best be found in the Hosmer and Lemeshow contingency; the segmentation of the sample into deciles builds the foundation for implementing predictive modeling into enrollment management practice. The review of the observed versus predicted probabilities of cases falling into each these deciles provides a likelihood ratio of enrollment for each category of scored cases.

The criteria of the study met with important results towards the advancement of the research questions. The introduction of the variables used within the logistic regression analysis provided enhancement to the probability of prediction of enrollment. This model generated through the analysis of the Developmental Group provided a foundation for the testing and application of the model in future iterations.

Though there was question to the actual fitness of the model, nonetheless actual improvement in the probability of prediction did occur. This model method increased the predictability of the enrollment in the subsequent entering class of 2010 and provides testament to the predictive tool in enrollment planning and study.

Residuals and Multicollinearity

In the course of logistic regression analysis in this study, the analysis of residuals was purposefully contained. Specific residuals, such as the predicted probability and predicted group membership were used in the study, while other residuals (such as the Cook's distance) were not addressed. In the case of evaluating probability of enrollment, the examination of residuals provides a portion of the variance seen in the outcome variable which is not described by the predictor variable (Vogt, 2005). Additionally, the residual analysis provides the researcher with the information necessary to isolate points for which the model fits poorly, and to isolate points that exert undue influence upon the model design (Field, 2005). The analysis of the first aspect of residual analysis (the point at which the model fits poorly) may have enhanced the fitness of this model design. The analyses of points which exert undue influence, however, are negated by the nature of the research design. To evaluate the probability of enrollment, then the whole variability of each case must be preserved since the likelihood of a subsequent sample having extreme outliers is fairly certain. Future study of model design in logistic regression analysis specifically for increasing probability of prediction in enrollment may benefit from residual analysis in order to develop a better fitness of the model.

Multicollinearity confounds the examination of multivariate regression analysis. Though primarily a linear regression consideration, logistic regression is performed in a manner similar to the linear regression design and can be complicated by multicollinearity conjunctions. Multicollinearity exists when two, or more, independent (predictor) variables correlate to such high degree that their correlation approaches a perfect collinear relationship, and are thus indistinguishable (Field, 2005). In this study, it is a possible consideration that the collinear effect of two or more variables could confuse the logistic regression and hamper the examination of the coefficients generated through model design. However, the sum exploration of the total sample's case variables is necessary for model development since the model relies on continuity of input across subsequent models. Maintaining the contributions of all variables across all cases derives the increase/decrease of predictability *en total*. The exclusion of predictor variables can strengthen the model; but by exclusion, there is a certain devaluation of particular case components (in this case, elements which are eliminated *a prior* so that the design becomes prejudicial, for example, of student backgrounds).

Predictive Enrollment Application

The information that develops from the logistic regression analysis of entering classes of freshmen provides valuable considerations for the enrollment management profession. As discussed earlier, logistic regression in enrollment management benefits an institution of higher education in multiple facets of the institutional operation. Anticipating and planning the enrollment of an incoming class of freshmen

allows institutions to make decisions about space and resource allocations necessary to meet the needs of the incoming class. Additionally, budgetary and resource distribution among the components of the institution benefit from the accurately predicted enrollment of incoming students by forecasting revenue production and identifying constituent areas of the institution which will need additional support. Conversely, these same techniques for prediction may be used with post-matriculant students at the institution. Applying logistic regression towards prediction of persistence and retention of students at the institution provides the same informative data to assist in the planning and resource allocation within the context of an operating institution and the needs of its constituents and students.

DesJardins (2002) introduced the idea for this study as an effective tool for recruitment and marketing to admitted students. In the recruitment of students to an institution of higher education, the "admitted" student is still far from the "committed" student, and continued communication and recruitment efforts persist even though a student has been admitted to the institution. In the context of this study, the distribution of individual cases (students) in the samples (enrolling classes) provides the opportunity to apply the logistic regression statistical power to application. As seen in the development of the logistic regression model, solving for the logistic regression equation for a given case is an indication of probability towards enrollment. By using the Hosmer and Lemeshow contingency table as a guide, the cases are distributed into equal deciles of categorization. Each of these deciles indicates a ten factor of probability from 0% to 100%. Each score from 0 to 10 then indicates a

classification synonymous to a given percentage of probability. Taking this into consideration, then each individual student case can be assigned a percentage of probability for enrollment.

Building upon this structure, each of the deciles becomes a score associated to the groups of students within these scores. A student with a score of 2, for example, would be unlikely to enroll into the institution. A student with a score of 8, however, may be much more likely to enroll in the institution. A score of zero and a score of 10 would almost always not enroll/enroll respectively. By scoring the cases, groups of students segment out by probability of enrollment. The enrollment management profession uses this information to build strategies for recruitment and communication. Students with high scores (e.g., 9 and 10) are almost assuredly going to enroll in the institution. These students receive the continued communication and connection to the institution through marketing materials and contact with recruiting staff, but the investment is made to maintain this level of enrollment probability and is not overdone. Those students below 5 are not necessarily worth investing additional effort other than standard marketing and limited communication from recruitment staff. The students scoring in the 6, 7, and 8 range are highly impacted by continued outreach and specific, strategic communication and marketing may be the additional element which will push these “fence sitters” over to the side of enrolling in the recruiting institution. Abstractly, the logistic regression curvilinear model provides a conceptual representation of this investment of resource. Students closer to the 100%

enrolled probability need only a little more investment of interaction to reach the level of matriculation to the institution.

Recommendations for Future Study

Regression analysis provides powerful interpretive tools for forecasting future events based upon existing data. The study of probability of enrollment with logistic regression analysis is particularly effective in predicting the possible outcomes of student behavior. Though this prediction never approaches 100% accuracy, the ability to make decisions with more significance than mere chance justifies the study and interpretation of this technique. This study alone did not provide the definitive model of prediction for enrolling students; there are potential improvements and further refinements which can aid this model.

The choice of predictor variables used in this study derived from the DesJardins study (2002) with the intent to replicate these predictors as closely as possible. The choice of predictors in that study were both good predictors evolved from solid research and theoretical backgrounds, but were not necessarily the ideal predictors for the study institution in this design. There are geographical differences and primary college entrance exam propensity which could influence the effectiveness of the chose predictors. Further iterations of this study should be conducted with the presence of the study institution first, the data of that the study institution collects about incoming students, and continuity of the collection of information.

The goodness-of-fit statistic in the logistic regression model failed to meet the threshold of validity (or rather, maintained significance rejecting the null hypothesis

that the probability of prediction was significantly different than the observed occurrence). Though the study still maintained validity across all other *post hoc* analyses, this fitness is an important indicator of effect and generalizability using the various predictors involved in this model. Residual investigation and analysis could provides a means of culling the predictor variables in the course of this design; though, a concern in restricting predictors could fall into a inadvertent prejudicial model with the inclusion of demographic predictors which may be indicative to a specific geographic location, socioeconomic status, and/or culture. Also, the various predictor variables themselves exclude from the model's predictive analysis individually (i.e., insignificant Wald scores). Further refinement of the chosen predictor variables will benefit the further evolution of an enrollment prediction model; these steps will enhance the fitness of the model by diminishing the effect of detracting predictors.

Conclusion

Statistical methodologies provide a wealth of interpretative and predictive tools for the study of the behavioral sciences. Logistic regression especially addresses the predictive nature of events involving human subjects by expressing the dichotomous values of events in a regression analysis akin to linear regression. The ability to predict behavior along the lines of dichotomous events provides researchers and practitioners alike with valuable information about the probabilities of events taking place. In enrollment management, there is a wealth of information which can be directed from this information as it becomes more accurate in predicting actual outcomes.

This study envisioned the adaptation of predictive modeling of human behavior of individual student's choice about attending an institution of higher education. Though it can never approach 100% accuracy (the human mind's decision making capacity involves infinite variables of various weights taken into consideration when making decisions), nonetheless the value of predicting behavior beyond chance fuels strategic planning on a multitude of levels when considering the actual students who will arrive at an institution the first days of an academic year.

The study sought the formation of a logistic regression model for the purposes of achieving just that axiom; predicting enrollment behavior of an incoming class. The development of a model was conducted with the logistic regression method, tested against a holdout sample, and implemented in a subsequent entering class to quantify the effectiveness of the design. The findings were of a significant level which can conclude some accountability of the variance found in the decision making process. Though small in comparison to all possible acting influences upon choice, the model provided evidence to suggest that the portion alone discovered through the model design would enhance the probability of predicting enrollment above chance alone.

Enrollment management in its various forms benefits from observation and prediction immensely in the operation of the profession to the benefit of the institutions of higher education it represents. Building better models of predictive enrollment strategy provide enhancement to the overall operation of an institution of higher education, a more responsive institute of higher education to its students, and a better steward (in the case of public institutions) of the resources the public provides

to the fulfillment of the educational mission of the institution. Logistic regression analysis is a definitive tool for all enrollment management professionals in the conduction of their work. Building better systems of enrollment management provokes better institutions of higher education to the benefit of future students in our society.

BIBLIOGRAPHY

American Association of Collegiate Registrars and Admissions Officers. (2002). *The 2002-2003 AACRAO Member Guide.*

Astin, A. (1965). *Who Goes Where to College?* Chicago: Science Research Associates.

Beale, A. V. (1970, November 15). The evolution of college admission requirements. *National Association of College Admissions Counseling* , 14-15.

Bowen, H. (1977). *Investing in learning: The individual and social value of American higher education.* San Francisco, CA: Jossey-Bass Publications.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review , 75*, 1-3.

Brubacher, J. S., & Rudy, W. (1997). *Higher education in transition: A history of American colleges and universities, 1636-1976* (3rd Edition ed.). New York: Harper & Row.

Cabrera, A. F. (1994). Logistic regression analysis in Higher Education: An applied perspective. In J. C. Smart (Ed.), *Higher Education: Handbook of Theory and Research* (Vol. 10, pp. 225-256). New York: Agathon Press.

Chapman, D. W. (1978). Improving information for student choice: the national effort. *National Association of College Admissions Counseling Journal , 23* (1), 25-26.

Council for Aid to Education. (1997). *Breaking the Social Contract: The Fiscal Crisis in Higher Education.* Santa Monica, CA: Rand Corporation.

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

Dey, E. L., & Astin, A. W. (1993). Statistical alternatives for studying college student retention: A comparative analysis of logit, pobit, and linear regression. *Research in Higher Education , 34* (5), 569-581.

Duffy, E. A., & Goldberg, I. (1998). *College admissions and financial aid, 1955-1994.* New Jersey: Princeton University Press.

Field, A. (2005). *Discovering Statistics Using SPSS.* Thousand Oaks, CA: SAGE Publications, Inc.

Fuller, W. C., Manski, C. F., & Wise, D. A. (1982). New evidence on the economic determinants of postsecondary school choice. *Journal of Human Resources , 17*, 477-498.

Gose, B. (1999, May 7). Colleges turn to consultants to shape the freshman class. *Chronicle of Higher Education , XLV*, p. 35.

Greene, W. (1993). *Econometric Analysis.* New York: Macmillan.

Hanushek, E. A., & Jackson, J. E. (1977). *Statistical Methods for Social Scientists.* San Diego, CA: Academic Press.

Hoenack, S. A., & Wieler, W. C. (1979). The demand for higher education and institutional enrollment forecasting. *Economic Inquiry , 17*, 89-113.

Holland, J. L. (1958). Student explanation of college choice and their relation to college populatiry, college productivity, and sex differences. *College and University , 33* (3), 312-20.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd Edition ed.). New York: John Wiley & Sons, Inc.

Hossler, D. (1984). *Enrollment Management: An Integrated Approach.* New York: College Entrance Examination Board.

Jackson, G. A. (1978). Financial aid and student enrollment. *Journal of Higher Education* (49), 548-574.

Kahler, D. (2008). Technology-Enhanced Recruitment Communications. In B. Lauren (Ed.), *The College Admissions Officer's Guide* (pp. 145-163). Washington, D.C.: American Association of Collegiate Registrars and Admissions Officers.

Kerr, C. (1990). The American mixture of higher education in perspective: Four dimensions. *Higher Education , 19* (1), 1-19.

Kinzie, J., Palmer, M., Hayek, J., Hossler, D., Jacob, S. A., & Cummings, H. (2004, September 1). Fifty years of college choices: Social, political, and institutional influences on the decision-making process. *New Agenda Series , V* (3), pp. 1-66.

Leslie, L., & Brinkman, P. (1988). *The economic value of higher education.* New York, NY: American Council on Education and Macmillan.

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

McPherson, M. S., & Shapiro, M. O. (1993). The search for morality in financial aid. *Academe , 79* (6), 23-25.

Miller, R. I. (1999). *Major American higher education issues and challenges in the 21st century.* London: Jessica Kingsley Publishers.

Noel-Levitz. (2007). *How to Use Pertinent Decision Data in Your Admissions Office to Enroll the Students You Want.* Coralville: Noel-Levitz, Inc.

Noel-Levitz. (2008). *Qualifying Enrollment Success: Maximizing Student Recruitment and Retention Through Predictive Modeling.* Coraliville: Noel-Levitz Inc.

Pascarella, E. T., & Terenzini, P. T. (1991). *How college affects students: Findings and insights from twenty years of research.* San Francisco, CA: Jossey-Bass Publishers.

Paulsen, M. B. (1990). *College choice: Understanding student enrollment behavior.* Washington, D.C.: Association for the Study of Higher Education.

Rudolph, F. (1990). *The American college & university.* Athens, GA: University of Georgia Press.

SAS Institute, Inc. (1995). *Logistic Regression Examples Using the SAS System. Version 6, First Edition.* Cary, NC: SAS Institute, Inc.

Vogt, W. P. (2005). *Dictionary of Statistics & Methodology: A Nontehnical Guide for the Social Sciences.* Thousand Oaks: Sage Publications, Inc.

Welki, A. M., & Novratil, F. M. (1987). The role of applicant's perceptions in the choice of college. *College and University , 62* (2), 147-60.

Wolanin, T. R. (1996). The history of TRIO: Three decades of success and counting.