Fall 2015

Intentional Recruiting: Using business intelligence, data mining, and predictive analytics to identify characteristics of those students who enroll, and graduate; in support of university enrollment management

Stephanie L. Harris
Central Washington University, harris14@uw.edu

Intentional Recruiting:

Using business intelligence, data mining, and predictive analytics to identify characteristics of those students who enroll, and graduate; in support of university enrollment management

---

A Thesis

Presented to

The Graduate Faculty

Central Washington University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

Information Technology and Administrative Management

---

by

Stephanie L. Harris

November 2015

CENTRAL WASHINGTON UNIVERSITY

Graduate Studies

We hereby approve the thesis of

Stephanie L. Harris

Candidate for the degree of Master of Science

APPROVED FOR THE GRADUATE FACULTY

_____ _____
Dr. Bob Lupton, Committee Chair

_____ _____
Dr. Natalie Lupton

_____ _____
Dr. Laura Portolese Dias

_____ _____
Dr. Kevin Archer
Dean of Graduate Studies

ABSTRACT

INTENTIONAL RECRUITING: USING BUSINESS INTELLIGENCE, DATA MINING, AND PREDICTIVE ANALYTICS TO IDENTIFY CHARACTERISTICS OF THOSE STUDENTS WHO ENROLL, AND GRADUATE; IN SUPPORT OF UNIVERSITY ENROLLMENT MANAGEMENT

by

Stephanie L. Harris

November 2015

Using business intelligence (BI) and archival data from a division II, public comprehensive, university in Washington State, the researcher identified specific characteristics of those students who enrolled, persisted and completed to undergraduate degree attainment. These characteristics created an applicant profile to be used in future enrollment management activities for intentional recruiting, while the predictive models for enrollment and completion inform administration to improve tuition revenue planning and budgeting, and to forecast future enrollment yield.

Keywords: Intentional Recruiting, Academic Analytics, Predictive Enrollment Model

ACKNOWLEDGMENTS

Thank you to Dr. James DePaepe who suggested the subject matter for this masters study when I was not sure what direction to take. Thank you Jim, for a solid beginning, for the opportunities provided, the encouragement, and the ongoing support.

Thank you to Dan Matthews, who I only half-jokingly say taught me how to build predictive models 30 seconds at a time. Who laughingly told me "I guess you have to start over" in what I'm sure he considered a teachable moment when I asked – but the model doesn't take into consideration x, y, and z. Who acted as a sounding board, a mentor, and a friend during the dark hours.

Thank you to Dr. Natalie Lupton for your extensive help in editing and turning what I'm sure you sometimes thought was a unorganized mess into something that was worthy of defending. Drafted into service unexpectedly, your input, and hours of editing and feedback have been immensely appreciated.

To the entire Information Technology and Administrative Management faculty and staff: I've said it before, but I want to go on record – the program is amazing and the instruction and curriculum applicable in so many ways unimagined at the beginning of this journey. I use things I learned in the program EVERY SINGLE DAY. Keep on keeping on!

And finally to my family and friends – Thank you for your support, love, and the occasional metaphorical head-slap to bring me back to reality. I love you.

TABLE OF CONTENTS

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

Chapter One

Introduction

Abstract

Using business intelligence (BI) and archival data from a division II, public comprehensive, university in Washington State, the researcher identified specific characteristics of those students who enrolled, persisted and completed to undergraduate degree attainment. These characteristics create an applicant profile to be used in future enrollment management activities for intentional recruiting, while the predictive models for enrollment and completion inform administration to improve tuition revenue planning and budgeting, and to forecast future enrollment yield.

Purpose of the Study

Universities routinely use student records to measure persistence rates and related research for their institutions. The purpose of this research is to, through the use of BI tools, data mining techniques, and predictive analytics, identify predictors for those students who will enroll and complete through degree attainment at the institution under study. These predictors may be used to facilitate effective student recruiting as well as for budget and planning purposes at the institution.

This study will provide the Division of Enrollment Management (DEM) actionable information. First, the results of this research may be used to intentionally recruit students identified by the study as more likely to persist to graduation. Doing so, the institution will recognize a significant return on investment (ROI) through increased

enrollment, retention and persistence to completion by those targeted students in conjunction with recruiting expenditure savings. Second, using the predictive models developed, the administration will purposefully plan additional resources for those populations of students, recruited and enrolled who need ancillary services to successfully persevere to completion.

This paper will present the results of a case study that determined the statistically significant characteristics of those students who enrolled, persisted and completed to degree from the post-secondary institution studied. After answering this descriptive question, the researcher developed a predictive model, which could be used at this and other similar institutions by administration when recruiting potential future students to identify those populations most likely to complete a course of study and graduate; and to project enrollment and persistence rates based on current and incoming populations.

Problem and Background

Problem statement. State and federal support of public institutions of higher education has decreased dramatically since the early 2000's. These institutions now rely heavily on tuition as a major source of funding. Enhanced, more precise recruiting of new, first-year students is vitally important to these institutions and improvements on the recruiting process add value to the institution. Administration's ability to more precisely and accurately predict student success supports planning and budgeting at the enterprise level.

Financial need. State and federal funding of public higher education institutions has dropped nationally by 24% over the last 15 years (GAO, 2014). This disinvestment has prompted many institutions to look for ways to improve operations and increase efficiencies where possible, and the importance of recruiting and retention has been recognized. A 2013 study found the median cost to institutions in the U.S. to recruit a single new student varies from an average of $457 at 4-year public institutions to $2,433 at 4-year private institutions (Noel-Levitz, 2013). By supplying enrollment management divisions with specific characteristics of those students who are most likely to enroll and successfully finish a course of study to completion thereby earning a degree, the cost of recruiting per student should be reduced thus freeing up funding for other needs.

Enrollment management. Enrollment management in the context of higher education is understood to bring together, under one division, the functions of admissions, student records and financial aid (Epstein, 2010). At the institution of this case study the DEM has a much more limited footprint, only the Office of Admissions reports directly to DEM. The Registrar's office (student records), Office of Financial Aid, and Student Success Division all work with DEM to support recruiting and retention efforts, however they are not direct reports to the enrollment management director. In addition, many areas and departments on campus have their own recruiting efforts, some of which work with DEM to achieve recruiting goals; however, most operate independently. Enrollment management should encompass three main themes: recruiting success, retention strategies, and promoting the institution through brand

awareness with marketing and advertising (J. Swiney, personal communication, July 20, 2015).

As indicated, DEM at the institution under study is responsible for the bulk of the institutions recruiting efforts. The recruiting process used has four distinct processes – search management, inquiry pool, applicant/admitted pool yield activities and personal touch activities. Starting with search management techniques, for incoming first year students, DEM purchases SAT scores limited to a geographical area and defined by test scores and GPA's that historically have produced students who attend. A strategic action plan and a communication plan are developed to meet the institutions recruiting goals (e.g. increase diversity, increased out of state or international students, etc.), which have been set by upper administration. The communication plan developed makes contact by email, telephone or letter based on the future students indicated interests and/or demographics. Those who respond requesting more information go into the inquiry pool (J. Swiney, personal communication, July 20, 2015).

Potential students in the inquiry pool are contacted and provided with additional resources and information to answer any questions; may be provided with faculty or student group contacts for their area of interest; and are encouraged to set up a visit and campus tour. They may receive a focus view book (a brochure on the institution that highlights specific areas) based on their area of interest, a letter from the chair of the department which houses their area of interest, and ongoing email correspondence from

DEM personnel. The goal is to move the potential student from the inquiry pool to applicant status (J. Swiney, personal communication, July 20, 2015).

Once the potential student has completed an application the institution makes a determination of acceptance within a few days; therefore, the applicant pool is actually the applicant/admitted pool. At this point the institution turns its focus on the applicant/admitted pool to yield activities. These activities include a spring open house inviting all students to campus with their families, orientation during the summer where new students will be enrolled in classes, see their dorms and learn more about campus life, and welcome weekend the beginning of fall quarter which coincides with housings move in date. All of these yield activities are designed with the purpose of converting potential students to enrolled students (J. Swiney, personal communication, July 20, 2015).

In concert with these activities, DEM has a fourth area of focus they term personal touch activities. Personal touch activities occur when recruiters go out into the high schools, college fairs, and attend (or offer) hosted events to tell the story of how this institution can meet the prospect's personal and educational goals; and to encourage, assist, and explain to students the resources available to help with applications, financial aid, student support services and anything else the potential student may have questions about. Recruiters from DEM attend approximately 1,000 events annually at the institution under study (J. Swiney, personal communication, July 20, 2015).

Reporting is another primary function of DEM. The director must keep the institution leadership informed as to the progress toward enrollment goals, update various departments as to how many students they should plan for, and provide to student services the current status of enrollment projection attainment, in a timely fashion. These reports provide information so that business decisions can be made such as: how many dorm rooms will be used; how many students will dining services be serving; and how many advisors will be needed for the various student groups. DEM leadership focuses on creating and providing projections and actual recruiting, enrollment, and attendance numbers to both upper administration and to ancillary departments across campus. Upper administration and Business and Financial Affairs require frequent data updates on tuition revenue and net tuition yield so that they in turn may plan and budget. Each new first year student recruited equals about $6,300 net in new tuition funding; overhead, building and capital projects, and financial aid equal approximately $2,700 to complete the $9,000 annual tuition cost (J. Swiney, personal communication, July 20, 2015).

Key Definitions:

A complete glossary of terms used in this research can be found in the appendices section of this paper (See Appendix A). Definitions of key terms are outlined here to provide the reader with context in the introduction section.

- ***Business Intelligence***: "is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance" (Gartner IT, n.d.)

- ***Completer***: "A Student who receives a degree, diploma, certificate, or other formal award. In order to be considered a completer, the degree/award must actually be conferred" (US Department of Education, National Center for Education Statistics [NCES], n.d.).

- ***Completion***: Status achieved upon successful conclusion of all requirements in a degree program. For the purpose of this paper Completion refers specifically to completion of requirements in an approved course of study, which qualifies the student to earn a degree.

- ***Data warehouse***: "Massive database serving as a centralized repository of all data generated by all departments and units of a large organization. Advanced data mining software is required to extract meaningful information from a data warehouse" (BusinessDictionary.com, n.d.)

- ***Enroll***: A student is considered enrolled at the institution if they are registered in at least one course for the term in question (Data Cookbook – Central Washington University [CWU], n.d.).

- ***Enrollment Management***: "An organizational concept and a systematic set of activities designed to enable education institutions to exert more influence over their student enrollments. Organized by strategic planning and supported by institutional research, enrollment management activities concern student college choice, transition to college, student attrition and retention, and student outcomes. These processes are studied to guide institutional practices in the areas of new student recruitment and financial aid, student support services, curriculum development and other academic areas that affect enrollments, student persistence and student outcomes." (Hossler & Bean, 2012).

- ***Enrollment yield***: The number of admitted students who actually enroll in at least one course. Usually presented as a ratio or percentage of the whole (Steinberg, 2010).

- ***First Year***: A matriculated student who is in his/her first year of attendance at the university, who have not attended (or attempted course credit at) another university after high school graduation. Being classified in this category is without regard to summer term credits and/or transfer credits earned through Running Start, Cornerstone or another dual-credit high-school/college program. This category of student includes admit types: FYR (First year), FYT (First year transfer), and IFY (International first year) (Data Cookbook – CWU, n.d.).

- ***Persist/Persistence***: "The act or fact of persisting" (Merriam-Webster, n.d.). For the purpose of this paper, the author differentiates between persistence and retention as follows: Persistence or to persist indicates that the student continues enrollment through completion of degree attainment; whereas, retention indicates that the student returned the following academic year – usually measured from fall to fall.

- ***Predictive Analytics***: Technology that learns from experience (data) to predict the behavior of individuals in order to drive better decisions (Siegle, 2013, p. 107). Involves extracting data from existing data sets with the goal of identifying trends and patters. These trends and patterns are then used to predict future outcomes and trends. While it's not an absolute science, predictive analytics does provide companies with the ability to reliably forecast future trends and behaviors (NG Data, n.d., para 1)

- ***Predictive Model***: "A mechanism that predicts a behavior of an individual, such as click, buy, lie, or die. It takes characteristics of the individual as input, and provides a predictive score as output. The higher the score, the more likely it is that the individual will exhibit the predicted behavior" (Siegel, 2013, p. 25).

- ***Recruit***: to attempt to enroll or enlist (a member, affiliate, student or the like) (recruit, n.d.).

- ***Retention***: Undergraduate degree seeking students who enroll consecutively from one academic year to the next academic year (Data Cookbook – CWU, n.d.).

- ***Student Success***: Provides educationally purposeful programs, events services and activities that promote academic, personal and professional growth within and beyond the classroom (Central Washington University, n.d.).

Research Questions

1. Can statistically significant characteristics be identified to provide a basis for intentional recruiting at the institution under study?

2. Using these characteristics, can successful completion of an undergraduate degree by a first year student be predicted?

3. What are the characteristics identified?
4. Are the characteristics different in populations that identify as diverse by race and/or ethnicity?

The researcher hypothesizes that targeted recruiting efforts based on recommendations will recognize a substantial ROI in the form of higher graduation rates, focused recruiting expenditures, and improved institution awareness and reputation by constituents.

Scope and Overview of the Study

This study analyzed descriptive data for those undergraduate degree-seeking students who were enrolled at the institution under study beginning Fall 2004 quarter through and including Spring 2015 quarter (n=7,077). First year students in that population who successfully completed their degree and graduated were identified and individual student characteristics generally known at or before application were determined. Predictive models to be applied across the entire student population were built.

The researcher analyzed and built predictive models for two populations. The first population contained all first year degree seeking undergraduate students who first enrolled in or after Fall 2004 quarter up to and through Summer 2009 quarter. By limiting the population thus, each student enrolled had six years to complete their degree. Federal financial aid limits eligibility for Direct Subsidized Loans to six years therefore,

six years (or 150%) is considered the standard generally maximum acceptable time frame for first year students to earn a first four-year degree (DOE, 2013). The second population studied is a sub-set of the first defined by the same initial parameters (all first year degree seeking undergraduate students who first enrolled in or after Fall 2004 quarter up to and through Summer 2009 quarter) with the additional limitation that they must be defined as a 'student of color'. The institution defines students of color as students who have self-reported their race as African American/Black, Alaskan/Native American, Asian, Hawaiian/Pacific Islander, Multiracial, or their ethnicity as Latino/Hispanic – or any combination of these. It is important to study the population students of color separately because the literature states in general those students who are successful at completing a degree are middle-class, white, females and that diverse populations have different characteristics and motivations than European/White students (Hanover Research 2011; Inman & Mays, 1999; Faircloth, Alcantar, & Stage, 2015; Cerna, Perez, & Saenz, 2009). To avoid reducing diversity on campus these differences must be addressed when recruiting.

The sample size for this study is 7,077 students who were enrolled in the University between the Fall 2004 and Spring 2015. A data warehouse was built in 2014 using the institutions transactional system of record as source data. Additional datasets introduced include census data, national clearing-house data, and GIS datasets. The data warehouse was queried using the OLAP WebFocus, a proprietary BI software, for the dataset of all undergraduate degree-seeking students who were enrolled Fall 2004 quarter

or after, through and including Spring 2015. The data set was cleaned using traditional data cleansing methods. These data were then analyzed using RStat, a WebFocus component built on the open source R statistical software language. Statistically significant characteristics were identified and predictive models were built after using data mining best practices. Predictive models for each population were developed with the separate analysis techniques of decision trees and logistic regression. The use of two different analytics methods validates and strengthens the results.

Limitations and Delimitations

The researcher designed this study to include only those predictive characteristics (independent variables) generally known about students when recruiting actions are being taken. While first quarter financial aid disbursed will not be specifically known about a potential recruit, this amount is derived by financial aid offices through formulas that incorporate socio-economic indicators, which are more commonly known at the time of recruiting and have been identified once enrolled for yield predictions. It is true, some prospective students may choose not to fill out and submit the free application for federal student aid (FAFSA) or may not be accurately identified as to their socio-economic status by census data; this is equally true of students within the population of the study and, therefore, has been accounted for in the predictive models.

One may question how a first year student who is defined as "A matriculated student who is in his/her first year of attendance at the university, who has not attended

(or attempted course credit at) another university after high school graduation" may have characteristics that include transfer credits and/or a GPA from another external institution (CWU, Datacookbook, n.d.). This category may include those students who completed running start or other similar college in the high school programs, as well as students who took advanced placement (AP) or international baccalaureate (IB) courses in high school and passed the AP or IB exams for that area with a high enough score to earn college credit. These students are identified by transcripts and lists, which can be bought prior to recruiting and are often identified when they are inquiring for additional information of an institution.

Potential students' motivation, intrinsic values, or religious affiliation are not easily ascertained prior to student recruiting and as such were not appropriate to, nor were they included in this study.

Importance

The benefit of this study will be to provide administration with data derived from statistical prediction models developed during this study; to be used both by DEM and by the Student Success Division to improve student recruiting and persistence and for added planning and budgeting of required resources to support these efforts. Using the student profile(s) provided, recruiters will be able to target those populations who traditionally enroll and complete at this and similar institutions.

This research may provide the impetus for similar institutions to adopt and implement BI solutions and to apply predictive analytics to improve operational efficiencies and ultimately to support student success. Additional benefits of this research may include publishing the statistical prediction models developed in this study as a contribution to research in higher education.

It is understood that selectively recruiting only those populations who are profiled as more likely to enroll and complete from the institution under study thereby reducing access to higher education within some diverse populaces is a risk, which must be guarded against. This is the reason for doing separate analysis and modeling on the subset students of color. If implemented, the models should be used in concert.

Chapter Two

Review of the Literature

Introduction

While the literature regarding enrollment management (EM) is extensive, the use of predictive analytics in EM is still fairly new. This literature review first examines why enrollment matters to institutions of higher education, how EM evolved in higher education and what the general functions and goals for EM are today. Business Intelligence, knowledge management, data mining, and analytics within the domain of higher education are defined, described, and supplied with concepts demonstrating their use in higher education. Information on how many and at what level institutions are using academic analytics (analytics in higher education), the primary issues preventing more institutions from adopting analytics, and what the next steps may be for academic analytics will be identified. The author presents four case study examples of data mining and predictive analytics in EM and then outlines the student characteristics found to be relevant in these studies pertaining to predicting enrollment by applicants. Finally the author provides an aggregated list of the characteristics, which have been shown to be statistically significant pertaining to student persistence and retention with the specific requirement that the student sample studied has enrolled, persisted, and completed their course of study to degree.

Enrollment Management Overview

Evolution of enrollment management. Post World War II, public institutions of higher education had a 'build it and they will come' mentality – and it worked. This changed in the 1970's. The combination of decreasing population, diminishing state and federal support, and the emergence of the for-profit sector changed the focus of admissions officers from gatekeepers to salesman (Kurz & Scannell, 2006). To adjust to this new paradigm, universities had to adopt new business practices where recruiting, marketing, and sales are must have proficiencies. Frank Campanella, then Executive Vice President of Boston College, coined the phrase *Enrollment Management* in a 1974 memo to his colleagues when he declared that Boston College must "bring together the previously siloed functions of admissions, student records, and financial aid" (Epstein, 2010, pp. 9-10). Campanella's ideas were not popular; however, the problems persisted and a year later the new role of Dean of Admission, Financial Aid, and Records was created with Jack (John) Maguire earning the position (Epstein, 2010). Maguire outlined his EM strategies in To the Organized Go the Students, published in the *Boston College Alumni Bridge Magazine*. They were: admissions should use marketing strategies; data matters; market analysis is essential; financial aid is a recruitment tool; and retention is an enrollment tool (Larson, 2013). Using these strategies Boston College became a thriving, select, national institution less than a decade after being on the edge of failure; with improved financial stability and increased enrollment figures (Epstein, 2010). These

same five strategies still drive much of the EM theory and practices today. Hossler and Bean (2012) went on to define enrollment management as:

> An organizational concept and a systematic set of activities designed to enable education institutions to exert more influence over their student enrollments. Organized by strategic planning and supported by institutional research, EM activities concern student college choice, transition to college, student attrition and retention, and student outcomes. These processes are studied to guide institutional practices in the areas of new student recruitment and financial aid, student support services, curriculum development and other academic areas that affect enrollments, student persistence and student outcomes (p. 5).

According to Kurz and Scannell (2006), the overarching end goal of EM is to, "… efficiently, effectively meet and exceed enrollment targets, especially net tuition revenue" (para. 14). To accomplish this, EM programs often address goals similar to the following:

- To coordinate staff, information flow and integrate decision making based on a complete picture of the student enrollment experience;
- To develop a global enterprise wide system with student data to facilitate research, planning, recruitment and communication;
- To develop a marketing plan targeting and attracting prospective, institution appropriate students;

- To implement tuition and financial aid strategies to attract and retain the diverse student body desired while still generating positive tuition net revenue;
- To analyze, develop and implement a plan to meet immediate and long-term student and institution demand; and
- To develop and implement an ongoing program of identifying, intervening, and supporting students at risk of not persisting.

Upon review of the above EM goals, it becomes apparent that the days of EM decisions based on the "build it and they will come" mentality are long gone.

Post-secondary funding in crisis. Over the last 15 years, the government has participated in large-scale disinvestment of higher education on both a state and federal level. In a December 2014 report, the US Government Accountability Office (GAO, 2014) reported public college revenue from all state sources decreased nationally by 12%. During this same time period, enrollment increased in these colleges by 20%. The result of this decrease in state support and increase in enrollment is a net decrease in state funding per student (nationally) of 24%. During the 2002-03 academic year, the median state funding per student nationally was $6,211 and in 2012-13 it was $4,695. Nationally, 2012 saw tuition become the primary source of funding, replacing state revenue sources for the first time in public higher education (GAO, 2014). For those students and colleges in the State of Washington the news is worse. State support has decreased 37.5% from fiscal year 2008 (FY08) to fiscal year 2013 (FY13) – the ninth highest decline nationally (Justice, 2013). In 2011, Washington State students began

paying more of their public education through tuition (52%) vs. state funding (48%) for the first time (Justice, 2011). The decline of state and federal support is forcing institutions to operate more efficiently and effectively. As the primary funding source, enrollment must be understood and managed in order for institutions to plan for the future. Data-driven decision-making**,** the ability to make informed decisions based on actual data, has never been more important for EM in higher education, and it has never been easier, especially for those with resources.

Business Intelligence

Knowledge management (KM). Data with an understanding of context is simply information. Once combined with experience and reasoning, information becomes knowledge upon which decisions can be based. Authors Serban and Luan (2002) suggest two definitions to relate this domain to higher education:

1. Knowledge management is about connecting people to people and people to information to create competitive advantage; and
2. Knowledge management is the systematic process of identifying, capturing, and transferring information and knowledge people can use to create, compete and improve (p. 1).

Knowledge management exists at the intersection of strategic management, information sciences, human resource management and information and technology systems management (Hutchinson & Quintas, 2008). Technology alone does not create

KM and KM is not just technology; however, without the advances in technology over the last 15 years, KM would be a theoretical argument rather than an actual implementable process. Business intelligence facilitates the operationalization of KM.

Business intelligence (BI). Defined by Gartner IT, BI is "… an umbrella term that includes the applications; infrastructure and tools; and best practices that enable access to and analysis of information to improve and optimize decisions and performance" (n.d.). These applications and tools include Online Analytical Processing (OLAP) and data mining software to be used against a data warehouse infrastructure. A 2014 Educause report indicates that BI reporting and data warehousing (defined below) are now the fifth and sixth (respectively) most rapidly changing core information technology systems in higher education. According to the 2012-13 Common Data Set, 80% of US higher education institutions have a BI reporting system and 71% have a data warehouse system. However, few of these systems facilitate KM because they are not structured in such a way that they can be used for analytics. Only 20% have BI reporting dashboards for analytics and 35% have a data warehouse for analytics. BI reporting dashboards to visually deliver the status of an institution's processes and activities is the first priority on the list of higher education's top-ten strategic technologies for 2014; while administrative/business performance analytics to target organizational resources and support organizational goals is sixth (Lang & Pirani, 2014).

Data warehouse. A working definition of a data warehouse is more easily understood to be a "massive database serving as a centralized repository of all data

generated by all departments and units of a large organization. Advanced data mining software is required to extract meaningful information from a data warehouse" (BusinessDictionary.com, n.d.). Data warehouses combine data from disparate sources to facilitate information and analytical processing, data mining, predictive and prescriptive analytics and reporting (Tutorials Point, n.d.).

Data warehousing was first referenced in 1990 by Inmon who defined it as a "subject-oriented, integrated, time-variant, and non-volatile collection of data that support management's decision making process" (Tutorials Point, n.d., p. 89). Subject oriented means a specific subject to be analyzed, for example student completion. Integrated refers to multiple data sources that provide a common view and definition, e.g., student completion will always refer to a student who completed a course of study and earned a degree. Time-variant means that historical data is kept in contrast to a transactional system where only the most recent data update is kept, e.g., if a student changes their major, records of both the first major and the current major are kept by date in the data warehouse. Non-volatile describes the fact that once the data record has been loaded into the warehouse it will not change, e.g., the first major will always be returned as the first major, it will never be over written (1keydata.com, n.d.).

Data mining. The Gartner Group (2000), an information technology research and advisory firm, defines data mining as, "the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and

mathematical techniques". In Data Mining and Its Applications in Higher Education, Luan (2002) further refines this definition for the institutional researcher domain as, "the purpose of uncovering hidden trends and patterns and making accuracy based predictions through higher levels of analytical sophistication" (p. 19). Data mining uses one of two basic constructs – unsupervised or supervised knowledge discovery. Unsupervised knowledge discovery is a bottom up approach that makes no assumptions and seeks to discover relationships within a data set. Supervised knowledge discovery seeks to explain those relationships once identified (Luan, 2002). Data mining tools predict behavior, find patterns, discover relationships and model future trends. These tools enable the analysis of historical data to create of actionable models.

Analytics

Determining the effect of decisions made, or the effectiveness of a given tool or scenario and developing predictive trends through the study of historical data is the definition of analytics (BusinessDictionary.com, n.d.). Data warehousing and data mining enable organizations to take large sets of their own data and apply predictive analytics in ways not imagined in the past. Predictive analytics uses data from the past to build models to predict the future, and forecast future events using cluster analysis, association analysis, multiple regression, logistic regression, decision trees, neural networks or text mining (Asllani, 2014). Predictive analytics may employ multiple models to achieve the desired outcomes. Analytics provide a means for an organization

to improve future performance using KM and historical data to identify those enhancements or changes most likely to support the identified goal. This author's research study outlined in this paper, used predictive analytics to create models that will predict whether or not any given student who enrolls will graduate.

Academic analytics. Fortune 500 businesses were among the first to use business intelligence, data mining and analytics. Their use generally focused on analytics for operations and marketing purposes. Applying BI principals to higher education, Zhao and Luan (2006) noted "Institutional data often contain valuable information essential for more in-depth understanding of students and their college experiences" (p. 7). A comparison of business analytics to academic analytics may look something like:

Table 1

Business Analytics to Academic Analytics Comparison

Audience	Business World	Higher Education
Current customers/students	Who are my most profitable customers?	Who are the students taking the most credit hours?
	Who are my loyal customers?	Who are the students likely to return for more classes?
	What customers are likely to defect to my rivals?	Who are the persisters at our university?
Potential customers/students	How do we convert shoppers to customers and close the sale?	What type of courses can we offer to attract more students?

(Luan, 2002)

Published in 2006, The Educause Center for Applied Research (ECAR) sought to identify if and how well institutions of higher education were using KM, information and technology. Defining academic analytics as the higher education counterpart to business

intelligence in the corporate sector, ECAR surveyed 1,473 institutions receiving 380 responses representing institutions from the US and Canada, with a broad range of enrollments, budgets, Carnegie classes, and support structures (Goldstein, 2005). Forty-six percent indicated they are reporting from their transactional system with no additional resources for academic analytics, while only 15% reported having a data warehouse with ETL and OLAP or dashboards. Those institutions with more advanced technologies and effective training reported higher levels of satisfaction and more active use with their systems. Only three percent identified their primary use as "what if" decision support, predictive modeling and simulation, or automated trigger of business processes (Goldstein, 2005).

The survey asked specifically about EM and found that advanced academic analytics were used in the EM recruitment function to:

- auto alert when enrollment metrics fall out of desired range;
- forecast future demand for courses;
- identify strongest potential admission prospects; and
- tailor recruiting strategies for a specific student.

Retention functions included:

- identifying academically at risk students, and
- alerting officials when academic intervention is warranted (Goldstein, 2005).

Goldstein (2005) summarized that using academic analytics was most successful in improving institutional decision making and meeting strategic institutional objectives.

Student retention and EM were found to be the business areas that reported the most success using academic analytics to improve outcomes in tuition yield, workforce productivity, admissions/EM results and student retention.

The 2012 study, Analytics in Higher Education Benefits, Barriers, Progress and Recommendations, delivers an update on the current state of analytics by providing benchmarks and identifying challenges or barriers encountered when adopting analytics in the higher education sector. While 69% of the 339 respondents indicated that analytics is a major priority at some level of their institutions, only 28% reported that this priority is enterprise wide. Data warehouse and BI systems are now being used 62% of the time to integrate, organize and summarize large data sets; however, most are still using frozen data. Eighty-four percent of the respondents indicated that analytics is more important to the success of higher education than it was two years ago, and 86% believe two years from now it will be more important than it is now. Most institutions who are using analytics are doing so to support EM, to monitor and guide student progress, and optimize resources (Bichsel, 2012).

In Building Organizational Capacity for Analytics, Norris and Baer (2013) specifically sought out institutions that showcased exemplary practices rather than the industry average, using a pool of institutions that had been recognized as such. Relying on analytics for competitive advantage, the for-profit and not-for-profit primarily online universities proved to be the most advanced in using predictive analytics to drive administrative and academic processes. Norris and Baer (2013) surveyed 40 of the

exemplary institutions and found virtually all were using some form of analytics for EM and to improve student success. Provided examples include the Virginia Community Colleges who are engaged on high school campuses to advise and recruit students and the University of Michigan who has improved student success through identification, monitoring, and support services targeting high risk students. Embedded predictive analytics trigger an action item targeting students with at-risk behaviors and tracking learner outcomes. The for-profit and online institutions lead the industry in this segment:

- American Public University System (APUS) developed a predictive model that is 91% accurate at predicting stop outs over the forthcoming five semesters, and review enrolled students weekly ranking them by likelihood of not being retained;
- Arizona State University retention rates have increased 4-5% using Sun Devil Tracking and eAdvisor;
- The University of Central Florida PhD-level data mining program has successfully identified 80-85% of their at-risk students;
- Purdue estimates retention in Signals-informed courses (those with predictive analytics applied) has improved by 20% and four-year degree completion rates by 4%;
- Rio Salado College, and University of Phoenix have developed predictive models to identify at-risk students and alert the appropriate administration (Norris & Baer, 2013).

As the use of analytics in higher education matures and stabilizes so will the attending terminology. Recently the term *Learning Analytics* or *Learning Academic Analytics* has started appearing in the literature (Van Barneveld, Arnold, & Campbell, 2012). This author understands Academic analytics to be defined as the use of business intelligence analytics processes in the academic or higher education sector to improve operations and efficiencies. Learning analytics are more narrowly focused on applying analytics processes to study and predict student learning outcomes, often using learning content management systems data. This study focuses on the domain of academic analytics.

Adoption challenges in higher education. Higher education traditionally lags behind business in embracing new technologies and adoption of using analytics is no different (Norris & Baer, 2013; Amburgey & Yi, 2011; Zhao & Lang, 2006). The primary reasons cited for this delay include lack of funding to adopt new technologies (Goldstein, 2005; Bichsel, 2012); lack of available analysts to perform analytics (Lang & Pirani, 2014; Goldstein, 2005; Bichsel, 2012; Rios-Aguilar, 2015); lack of consistency in data sets and data definitions (Bichsel, 2012; Faircloth, Alcantar, Stage, 2015; Rios-Aguilar, 2015); and lack of institutional wide support and/or failure to provide change management/training (Goldstein, 2005; Bichsel, 2012). Systemically when predictive analytics are not an enterprise wide priority, a failure to connect student data to student outcomes breeds a laissez-faire culture (Bichsel, 2012; Rios-Aguilar, 2015).

Moreover, organizational leaders tend to overestimate their institutions capabilities with data, information and analytics and underestimate the challenge of changing the culture to fully embrace and utilize embedded deployment of predictive and performance analytics. Norris and Baer (2013) conclude this sector faces significant challenges lacking the professional development, capacity building, and understanding the application of analytics by institutional leadership and practitioners at all levels.

The future of academic analytics. Academic analytics will continue to develop and grow as institutions expand their awareness of the possibilities and return on investment (ROI). Those institutions that view data as an asset and analytics as an investment rather than an expense are more likely to have academic analytics a priority enterprise wide.

Dashboards visually displaying performance metrics in near-real time are high priority for many institutions (Lang & Pirani, 2014; Rios-Aguilar, 2015). Combining data from disparate sources (traditional system databases, social media, near field RFID technologies) for added analysis is becoming more common (Bichsel, 2012; Lang & Pirani, 2014). Norris and Baer (2013) reported that most of the exemplary example institutions they surveyed were currently engaged in "large-scale, longitudinal data analysis and comparative research to discover insights into 'what works' in making students successful" (p. 27). This type of analytics and research will become more common as the tools, resources, and experience become more widely accepted and understood throughout the sector.

The next level of predictive analytics in higher education will include cross-institutional data mining. The Predictive Analytics Reporting (PAR) project is creating a data set of six institutions (American Public University System, Colorado Community College System, Rio Salado College, University of Hawaii System, University of Illinois-Springfield, and the University of Phoenix) containing almost 800,000 student records with oversight from the Western Interstate Commission for Higher Education (WICHE). Descriptive, inferential and predictive analytical tests will be applied to these de-identified student records with an ultimate goal of finding solutions to decrease student loss and increase momentum and success (Norris & Baer, 2013). Norris and Baer (2013) suggest future studies will likely include learning with workforce elements to identify those behaviors and experiences most successful in transferring into the workforce.

Academic Analytics Case Studies

Specifically looking at the EM domain this author found several case studies where analytics had been utilized to answer questions, address inefficiencies and improve processes in EM. The findings from four of these studies are outlined here.

Case study 1: Chen. Dr. Chau-Kuang Chen (2008) completed a three phase, longitudinal data study analyzing student enrollment records from 1962 to 2004 at Oklahoma State University (OSU). Phase one developed an autoregressive integrated moving average (ARIMA) model that when applied to the data returned a R^2 value of .96.

Phase two applied linear regression to the same data set with a R^2 of .97. Phase three compared the results of the two models concluding that the comparison was statistically insignificant and either model could be applied (Chen, 2008). The models addressed 15 variables and found that OSU enrollment is significantly and positively associated with two characteristics – Oklahoma high school graduates and one year lagged OSU enrollments (Chen, 2008). This study was one of the first to longitudinally analyze data across such a broad time frame in higher education and was cited in many of the data mining studies that followed in this area.

Case study 2: Amburgey and Yi. Using Decision Tree Analysis (ASE .379), Multiple Regression Analysis (ASE .382) and Neural Network Analysis (ASE .373) Amburgey and Yi (2011) analyzed 2006-2008 first year fall enrolling students (n=3576) at Saint Joseph's University to determine their end of first year GPA. The three models were compared against each other to determine the best fit using the Average Squared Error (ASE) with 11 independent variables. The models were found to be so similar that although the neural network would be the model of choice, any of the three could be used. The authors suggest that understanding a students predicted success allows EM to tailor communications based on the individual (honors information to a high achieving student, academic mentoring and student success resources to a marginal candidate) in addition to answering the obvious question of should they be offered enrollment (Amburgey & Yi, 2011).

Case study 3: Chang. Using applicant data from a large state university to predict enrollment behaviors of admitted applicants, Chang (2006) sought to identify if they enrolled randomly without significant and identifiable patterns, if certain types or groups of admitted applicants were more likely to enroll, and how well future enrollment could be predicted based on identification of these patterns. The population studied, admitted undergraduate degree-seeking freshmen, were analyzed using three predictive models: C&RT (decision trees), neural networks and logistic regression with a finding that where there was an agreement between the three models (66%) enrollment could be predicted 82% of the time when compared against actual enrollments. Overall the neural network models performed better than logistic regression and all three models performed better than the baseline (intuition). This study informed EM through knowledge of applicant's potential decision therefore communications were tailored to individuals; recruitment budgets and activities were redirected toward targeted populations; and academic programs were able to plan for desired degree majors (Chang, 2006).

Case study 4: Antons and Maltz. Historically Willamette University hired an outside consultant to estimate total enrollment yield using traditional statistical models resulting in estimates that varied from actual yield rates by as much as -16.9% to + 21.1% leading to "imprecise estimates of both yield and discount rate and ultimately significant declines in actual revenue accruing from tuition" (Antons & Maltz, 2006, pp. 70-79). In 2000 a partnership between their EM departments and a masters program with a data-mining component was created to develop models to 1) more accurately reflect actual

yield rates using Willamette's own data, and 2) expand understanding of the applicants positive decision factors. The model developed predicted yield rates with a variance of only 2.5%. The model enabled EM to apply varying financial aid packages at the student level to gauge the probability of the student enrolling based on distribution of loans and grants, and to provide much more accurate view of expected revenues at an institution level for the incoming class (Antons & Maltz, 2006).

Enrollment, Persistence and Retention

Businesses understand it is cheaper to keep a current customer than it is to find a new one. In the business world this term is referred to as customer churn and the amount of customers who sever relations with a business or company during any given time is defined as churn rate (churn-rate.com, n.d.). By the very nature of the business, higher education has a natural churn rate – all customers for one reason or another sever ties with a university, either because they have completed their course of study and graduate or because they are stopping or dropping out.

Identification of the characteristics of the population(s) who persist to graduation will inform an institution's EM offices and provide additional information for more targeted marketing efforts as well as increased accuracy in tuition yield modeling to support enterprise level planning and budgeting. The literature in the area of persistence and retention is extensive and this is not an exhaustive review but rather a touchstone to validate this study's findings with previous studies findings.

Significant Characteristics

Predicting enrollment. The student characteristics found to be statistically important when predicting enrollment were found to be: high school or previous college GPA (Amburgey & Yi, 2011; Chang, 2006; Antons & Maltz, 2006; Luan, 2002); high school rank (Chang, 2006); high school size / quality (Amburgey, & Yi, 2011; Chang, 2006); SAT or ACT scores (Amburgey & Yi, 2011; Chang, 2006; Antons & Maltz, 2006); admissions score index (Chang, 2006; Antons & Maltz, 2006); gender (Amburgey & Yi, 2011; Chang, 2006; Antons & Maltz, 2006; Luan, 2002); race/ethnicity (Chang, 2006; Antons & Maltz, 2006; Luan, 2002); age (Chang, 2006); religion (Amburgey & Yi, 2011; Chang, 2006); point of origin (Amburgey & Yi, 2011; Chang, 2006; Antons & Maltz, 2006); financial aid offered/package type (Antons & Maltz, 2006; Nandeshwar & Chaudhari, 2009); admission type (Chang, 2006); major program (Amburgey & Yi, 2011; Chang, 2006); frequency of pre-enrollment communications (Chang, 2006); source of initial contact (Chang, 2006); and/or non-basic skills courses taken (Luan, 2002).

Predicting persistence and retention. When aggregating students at the highest level studies show high school GPA (Lotkowski, Robbins, & Noeth, 2004; Johnson, 2008; Hanover Research, 2011); socioeconomic status (Adelman, 2006; Lotkowski, Robbins, & Noeth, 2004; Johnson, 2008); academic rigor of high school, post-secondary GPA (if any) (Adelman, 2006; Johnson, 2008; Hanover Research, 2011); race/ethnicity (Ice et al. 2012; Hanover Research, 2011); ACT/SAT scores (Lotkowski, Robbins, & Noeth, 2004; Hanover Research, 2011); advanced mathematics coursework in high

school (Adelman, 2006; Ice et al., 2012); proximity to the institution (Lotkowski, Robbins, & Noeth, 2004; Hanover Research, 2011); financial aid amount and ratio of loans (Johnson, 2008; Hanover Research, 2011); and first generation status (Johnson, 2008; Hanover Research, 2011) are student characteristics that have been found to be significant when studying retention and persistence.

Overview of significant characteristics. For ease of reference an overview of the significant characteristics for predicting enrollment and for predicting persistence and retention as identified by review of the literature is presented here:

Table 2

Significant Characteristics per Literature Review

Predicting Enrollment	Predicting Persistence and Retention
• High school or prev. college GPA • High school rank • High school size/quality • SAT/ACT scores • Admissions score index • Gender • Race/ethnicity • Age • Religion • Point of origin • Financial aid offered • Admission type • Major/program • Frequency of pre-enrollment communications • Source of initial contact • Non-basic skills courses taken	• High school GPA • Socioeconomic status • Academic rigor of high school • Post-secondary GPA • Race/ethnicity • SAT/ACT scores • Advanced mathematics coursework in high school • Proximity to the institution • Financial aid and ratio of loans • First generation status

These characteristics were used by in this study as a barometer that helped guide the researcher during the iterative process of developing the predictive models.

Summary of Literature Found

This literature provided a high level understanding of the history and relevant facts of EM, BI, and the use of data mining and analytics in the higher education sector. Also, the review supplied a list of student characteristics found to be significant in similar studies. This research is needed because the state and federal government have participated in a long-term divestment of higher education. Institutions have found to remain solvent they must better manage their operations. EM was developed precisely to answer this need.

Practicing KM principals, some institutions use existing data to make informed decisions. By understanding that knowledge is data with context applied, and when combined with experience and analysis, becomes the basis of data-driven decision-making. The literature provided an understanding of the evolution of data mining and analytics as pertains to higher education.

Analytics applied to the higher education sector, otherwise known as academic analytics have been used in a number of different ways. This review identified questions that analytics seek to answer; explored where institutions are in the process of moving from a transactional only system to one that can be used for analytics; how many institutions are currently using analytics; and the primary barriers identified by those

institutions who are not. Potential next steps for academic analytics have been identified. Example uses were provided in the form of identification of existing application at public and private institutions. Four case studies using analytics specifically for EM in higher education were summarized.

This study sought to identify an applicant profile to describe those students who have proven to be successful in completing a degree and graduating at a division II, comprehensive, public, regional university in Washington State. A list of characteristics found to be significant in similar studies was identified. The list was compared with the results of this research as a form of validation. It is not expected that all characteristics would align, but the literature shows a common set of variables that have been identified for potential inclusion.

Chapter Three

Study Design and Methodology

Study Design

This study sought to identify the specific characteristics of students who enrolled, persisted and completed to degree attainment at a division II, comprehensive, public university in Washington State with an annual enrollment of approximately 12,000 students. The research was completed on existing archival student data pulled from the system of record at the institution studied. The researcher utilized BI systems, data mining and predictive analytics to identify unique profiles of student populations for intentional recruiting and predictive models that, when applied to future applicants, will help to inform administration with forecasts of student enrollment and completion for planning and budgeting purposes.

The researcher obtained appropriate Institutional Review Board/Human Subjects Review Committee approval to use the data studied as it used existing archival data, which was then anonymized by removing student identification numbers and replacing them with non-identifiable research numbers.

At the institution studied, PeopleSoft, an Oracle based software, is the transactional system of record. PeopleSoft was implemented on the campus in 2004. As designed, historical reporting out of PeopleSoft is problematic as there is no system wide manner of tracking record changes. In 2014, a data warehouse was built and became the

system of record for historical reporting at this institution. An extract, transform and load (ETL) process is run nightly to populate the data warehouse from PeopleSoft. Within the data warehouse all records are date stamped and a historical record is maintained.

The Department of Institutional Effectiveness (IE) within the university studied builds, maintains, and reports from the data warehouse. Reports include required state and federal reporting for funding and financial aid as well as institutional accreditation, program accreditation, strategic planning, legislative data requests, and ad hoc data requests. Visual dashboards have been and are being developed to answer reoccurring questions and to support administrative and operational efficiencies.

The population under study is any first year student who first enrolled during or after Fall 2004 quarter through Summer 2009 quarter. This encompasses 7,077 students. The data warehouse was queried, the subject data set pulled, exported to Excel, cleaned, imported back into the BI application and analyzed using RStat (a statistical software based on the open source package Rattle written using the language R) where predictive models were built using decision trees and logistic regression. Using BI best practices the data set was split into training and testing sub-sets by the software. The predictive models that were built were then validated in Excel using the entire data set to verify accuracy and to ensure overlearning did not occur.

Data Preparation and Cleansing

WebFocus, a proprietary OLAP or information retrieval tool used in BI, was used to pull student records for the subject population from the data warehouse with the following identifiers: student ID (converted to a research ID), date of birth, original admit term, degree term code, age in term stopped out, gender, veteran flag, first generation flag, disability flag, race, ethnicity, WA resident, original postal zip code median income based on 2010 census, highest ACT, highest SAT, previous external institution type, max-degree code from external institution, transfer test or other credits taken, original admit term code, original admit type code, service campus code, developmental English flag, developmental Math flag, prior EASL flag, financial aid disbursed, degree term code (term degree was earned). These identifiers were chosen because they are generally known characteristics about a student when they apply and can therefore be used both to target recruiting efforts and to predict enrollment yield. The data was then exported to Excel for cleansing.

In Excel additional fields from existing data were created for use during analysis. To be able to use age at the time of admittance, the date of birth (DOB) field was split into DOB year, DOB month fields, and the admit term code was used to create new fields of admit term year and admit term month. From these four new fields *age_at_original_admit_term* was calculated ((admit term year – DOB year) + (admit term month – DOB month)/12). The researcher felt it was important to look at all aspects of race and ethnicity in this study; therefore, race and ethnicity were pulled as separate fields from

the data warehouse. However, for the purposes of state and federal reporting the following formula is used: if ethnicity equals Latino/ Hispanic then Latino/ Hispanic, else race; therefore a new field was created and named *Race_Ethnicity* using this formula and included in the analysis to ensure that all aspects of race and ethnicity were analyzed. Some prospective students choose to take the SAT, some the ACT, and some both exams. To effectively compare those students who took the ACT to those who took the SAT tests a new field, *ACT_to_SAT_Conversion*, was created based on the look up table ACT to SAT score concordance (ACT, 2013). This field was then used to create the field *highest_SATACT_composite* to compare the highest ACT converted score to the highest SAT score and report only this one score to be used in analysis. The spreadsheet was saved as a .csv format to be loaded back into WebFocus to be analyzed. Predictive models were built with RStat, the WebFocus component built on the Rattle platform using the open source statistical software language of R. This same data set was then cleansed again by filtering to remove all students who self-reported as European/Middle Eastern/White and not Latino/Hispanic; European/Middle Eastern/White and not reported; not reported and not Latino/Hispanic; not reported and not reported. This dataset was saved as the student_of_color data and saved as a .csv format to be loaded back into WebFocus and analyzed for the second part of the study.

Data Mining

Decision trees. Predictive models "…share the same objective: They consider the various factors of an individual in order to derive a single predictive score for that individual. This score is then used to drive an organizational decision, guiding which action to take" (Siegle, 2013, p 27). To conduct the data analysis and build the predictive models for this study, the researcher built decision tree and logistic regression models. Decision trees and logistic regression are used for classification purposes, and as such are appropriate for use in this study where the researcher is looking to classify students into clusters to answer the question: *What are the characteristics of a student who will complete a degree at this institution?* The resulting decision tree and logistic regression models answer this question by not only providing a description of the student profile but by also assigning each individual a predictive score. Ultimately each model answers the question with the statement: *A student with xx characteristics has a yy probability of completing a degree at this institution.* This predictive score will inform EM as to what action to take with any given candidate (Siegle, 2013).

These models are built using machine learning, which analyzes existing data to discover patterns that explain the data. The predictive modeling software, in this case RStat, derives the model through an act of trend spotting by exploring the data's broad range of factors. The modeling process must be done on data that includes both positive (did complete) and negative (did not complete) examples in the data for training (Siegle,

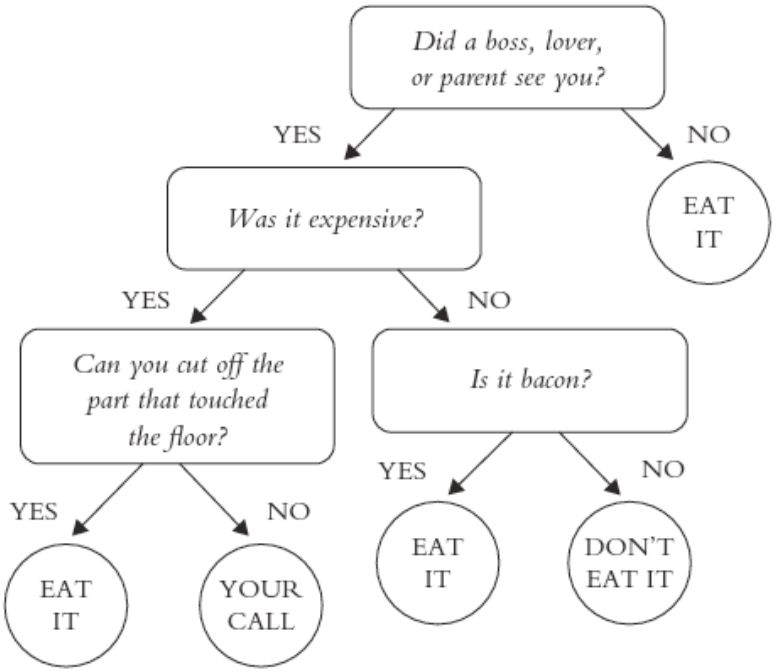
2013). Analysis with a particular goal or target (did/did not complete) is called supervised learning in data mining (Donalek, 2011).

To build a predictive model the data is split into two data sets. Training data is used to generate the predictive model and testing data is used to evaluate the predictive model. The testing data set must be quarantined from the training set which allows the model, once built, to validate itself and to provide a true evaluation of its ability to predict (Siegle, 2013). For the purpose of this study RStat has been set to randomly split 70% of the data set as training data and 30% of the data set as testing data.

A nonparametric classification method, a decision tree presents as a simple set of rules – if xx, then yy, else zz providing probabilities for each yy and zz classification. Described by Donalek, a decision tree (2011):

> is a tree-shaped structure that is derived from the data to represent sets of decisions that result in various outcomes – the tree's various end points. Decision trees do not require any assumptions on the distribution of the variables therefore they are considered nonparametric.

Once developed, a decision tree will predict the outcome when presented with new data – such as a new incoming student (Brown, Dehayes, Hoffer, Martin, & Perkins, 2012). Figure 1 shows a simple decision tree example designed to help you decide what to do about the lunch you just dropped – If your boss saw you drop it (no), then eat it, else was it expensive (no), then is it bacon (yes) then eat it, else don't eat it.



(Siegle, 2013, location 3038)

Figure 1: Sample Decision Tree

Decision trees risk overlearning or over-fitting, a scenario where the model creates a tree leaf for every individual in the data. This defeats the purpose of clustering characteristics into similar groups and is called overlearning. To prevent overlearning in decision trees the researcher must compare the results of the test data against that of the training data and if the test data indicates that overlearning has occurred, i.e. if the model predicts better in the training dataset than in the test dataset, the researcher must prune the

tree back to a point where the predictive results in the test data are very similar to the predictive results in the training data (Siegle, 2013).

For this study, the researcher uploaded the .csv file created as outlined data preparation and cleansing section into the RStat component of WebFocus to begin modeling. STUDENT_ID (the randomized student ID) was set as the identifier and DEGREE_COMPLETION_FLAG was set as the target and all other fields were left as input. The researcher then undertook an iterative process to build the model which consisted of: run the decision tree, check the overall error of the model, adjust the *complexity*, the *max depth*, the *min split*, the *min bucket* and the inputs one at a time each time checking the error of the model to identify if the change had made the model more or less accurate. Complexity refers to the complexity parameter (cp), the final complexity setting was .004. The cp controls the size of the decision tree, the smaller the cp the more complex the tree (Williams, 2010). Max depth refers to the number of levels of splits in the model. A maximum depth of 10 would mean that there are no more than nine nodes (branch splits). The final max depth was 30. Min split refers to the minimum number of instances in the node allowed. A min split of 20 means that there must be at least 20 students in a node. For this study the min split was 20. Min bucket refers to the minimum number of instances in the leaf resulting from the split. For example, if there are 21 students in a node (more than the min split level) and the split will create a leaf with seven students and a leaf with 14 students this would be allowed by the model; however, if the split would create a leaf with six students and a leaf with 16 students in

the corresponding nodes, this would not be allowed by the model. The final min bucket was set at seven. The inputs are the student characteristics (race, ethnicity, age, etc.) that will describe the student when the model is built. When no additional improvement could be made on the overall error of the decision tree, the model was saved. The researcher then used the variables the decision tree had identified as significant to provide a base understanding to develop the logistic regression.

Logistic regression. Logistic regression is "a class of regression where the independent variable is used to predict the dependent variable" (Statistics Solutions, 2015). The target in this study is a binary or dichotomous dependent variable – the student either did or did not complete their program to degree. Binary logistic regression applies statistical analysis to determine how much variance, if any, is explained on the dependent variable by the independent variables (Statistics Solutions, 2015). Logistic regression provides the probability of an event occurring unlike linear regression that strives to predict the change in the dependent variable based on the change in the independent variable (University of Strathclyde, n.d.).

In this study, the researcher used logistic regression to build a model to predict the probability that a given student with a particular set of characteristics will complete their program to degree attainment. Once the predictive model has been built, it can be extrapolated out to answer various tuition/enrollment yield questions that determine the yield based on the characteristics of the students who have enrolled. In addition, intentional recruiting can be undertaken based on the knowledge that students with an

identified grouping of characteristics are more likely to enroll and complete to degree at this institution and what that probability of success will be for any given grouping (cluster).

For this study, the researcher uploaded the .csv file created as outlined in the data cleansing section of this paper into the RStat component of WebFocus to begin modeling. STUDENT_ID (the randomized student ID) was set as the identifier and DEGREE_COMPLETION_FLAG was set as the target and all other fields were left as input. The researcher then undertook an iterative process to build the model which consisted of: run the logistic regression, check the overall error of the model, remove an input not shown by the logistic regression to be statistically significant one at a time each time checking the error of the model to identify if the change had made the model more or less accurate. If the change made the model more accurate continue, otherwise undo the last step and continue on to remove a different input. Initially the researcher used the characteristics identified by the literature review to guide which inputs to remove or retain. The goal of the process was to have included in the model only those characteristics, which were identified as being statistically significant. When no additional improvement could be made on the overall error of the logistic regression, the model was saved.

Applying the Models

Applying the predictive models built to the individual students in the studied population was done in the Excel spreadsheet(s) that were created in the data cleansing steps. These models assign a prediction and a probability of completion to each student based on their individual characteristics. The decision tree model clusters students and assigns a likelihood of completion to that cluster (node). The logistic regression assigns a value to each statistically significant characteristic, or independent variable, that the researcher used to create a formula to determine the probability that any given student would complete to degree. Using the Excel spreadsheet that was created in the data preparation and cleansing step, four new columns were created to apply the predictive models to the student population – DT_predictor, Decision_Tree_Predicted, Decision_Tree_Node, and Logistic_Regression. Each of these columns used the predictive models to determine the probability of the student in that row completing to degree. Where:

- E2 is a binomial first generation flag (1 yes, 0 no)
- F2 is the high school GPA
- G2 is the maximum GPA from an external institution
- H2 is the number of transfer credits accepted
- I2 is the number of transfer test and other credits accepted (which would include all transfer credits taken plus any AP or other credits earned through testing)

- J2 is a binomial flag to indicate if the student had to take a developmental math course upon entry to the institution (1 yes, 0 no)
- K2 is the first quarter financial aid disbursed amount. This amount will be known about any incoming students who have filled out the FAFSA and applied to financial aid and can be extrapolated out to be used as a socio-economic indicator when recruiting using current census data
- L2 is a binomial Race_Ethnicity_Asian flag (1 yes Asian, 0 no)
- M2 is a binomial Race_Ehtnicity_White flag (1 yes White, 0 no)
- N2 is a binomial Race_Ethnicity_LatHisp flag (1 yes Latino/Hispanic, 0 no)
- O2 is a binomial gender flag (1 Female, 0 male or unreported)
- P2 is a binomial admit type flag for first year students with no transfer credits - FYR, (1 yes, 0 no)
- Q2 is a binomial admit type flag for first year students with transfer credits – FYT (1 yes, 0 no)

To compare the decision tree prediction outcomes to the actual outcomes the researcher used a three-step process:

1. Using the first year all decision tree rules (See Appendix C) the Excel formula

   =IF(F2>=3.311,3,IF(AND(F2<3.311,J2=0,K2<1556,O2<0.5,F2>=2.792,G2>=3.21),91,IF(AND(F2<3.311,J2=0,K2<1556,O2<0.5,F2>=2.792,G2<3.21,M2<0.5,N2>=0.5),361,IF(AND(F2<3.311,J2=0,K2<1556,O2>=0.5,E2=0),47,IF(AND(F2<3.311,J2=0,K2<1556,O2<0.5,F2>=2.792,G2<3.21,M2>=0.5),181,IF(AND(F2<3.311,J2=0,K2<1556,O2<0.5,F2>=2.792,G2<3.21,M2<0.5,N2<0.5),360,IF(AND(F2<3.311,J2=0,K2<1556,O2<0.5,F2<2.792),44,IF(AND(F2<3.311,J2=0,K2>=1556),10,IF(AND(F2<3.311,J2=0,K2<1556,O2>=0.5,E2=1),46,IF(AND(F2<3.311,J2=1),4,"OTHER")))))))))

   was created and column C(2) populated. This formula assigned the decision tree node identified by the decision tree model to each student in the study population in the corresponding field of the new column Decision_Tree_Node.
2. The Excel formula =IF(OR(C2=3,C2=91,C2=361,C2=47,C2=181),"Yes", "No") was designed to interpret the decision tree node identified in step one and show the prediction as a yes/no response which indicates whether the decision tree predicted that student will complete their degree (yes) or not (no). The output of the formula is listed in the new column B(2) Decision_Tree_Predicted.

3. The Excel formula

   =IF(C2=3,0.69,IF(C2=91,0.67,IF(C2=361,0.64,IF(C2=47,0.6,IF(C2=181,0.54,IF(C2=360,0.42,IF(C2=44,0.39,IF(C2=10,0.36,IF(C2=46,0.36,IF(C2=4,0,"other")))))))))

   was designed to indicate the number of students in that node or group who are predicted to complete. This is read as 69% of the students who are assigned to decision tree node three should complete to degree. The output of this formula is listed in the new column A(2) DT_predicted.

The resulting columns allowed the researcher to compare the predicted outcomes to the actual outcomes as shown in the results section of this paper.

To compare the logistic regression model outcomes to the actual outcomes the researcher developed this Excel formula:

=0.27034295-0.68900301*E2+0.58708366*F2+0.11564206*G2-0.02217335*H2+0.02265524*I2-16.16388491*J2-0.00011885*K2+0.5045654*L2+0.5994673*M2+0.73342374*N2+0.33087185*O2-2.37225413*P2-2.412001*Q2

This formula creates an individual score for each student based on the characteristics found to be statistically significant by the logistic regression. These scores are shown in the new column D(2) Logistic_Regression and saved as recruitmodeldoball_firstyear_2015.7.15_satclean_wpredictors.xlsx. These same

procedures were applied to the dataset for students of color and saved as recruitmodeldoball_firstyear_2015.7.16_satclean_studofcolor_wPredictors.xlsx.

Evaluating the Models

Model accuracy as measured by the overall error provides the researcher with information as to the fit of the model however it may not be the best measure to evaluate the models developed for this study. When looking at the purpose of the study – to identify those prospective students who are most likely to graduate and therefore a) intentionally recruit them and b) develop tuition yield models; model validity is a better metric to use to determine model accuracy. Validity is the extent to which the model accurately predicts what it is designed to predict. Validity is measured by sensitivity and specificity (Parikh, Mathai, Parikh, Sekhar, & Thomas, 2008). Sensitivity is the probability that the model correctly predicted those students who did complete to graduation (yes/yes) or the true positive rate. Specificity is the probability that that the model correctly predicted those students who did not complete to graduation (no/no) or the true negative rate (Medcalc.org, 2015).

The predictive models when evaluated in the software produce a conventional two-by-two (2 x 2) table to display the results. An example of the output for the results of this study is shown in Figure 2.

Total Population	Predicted No	Predicted Yes
Actual No	TN	FP
Actual Yes	FN	TP

Figure 2: Conventional Two-by-Two Results Table

Where TP represents the true positives (predicted yes to graduate and did graduate), TN the true negatives (predicted not to graduate and did not graduate), FP the false positives (predicted to graduate and did not graduate) and FN the false negatives (predicted not to graduate but did graduate). Sensitivity therefore would be determined by TP (true positive) / TP+FN (true positive + false negative). Specificity would be determined by TN (true negative) / TN+FP (true negative + false positive) (Parkikh, Mathai, Parikh, Sekhar, & Thomas, 2008). The models were evaluated using sensitivity and specificity attainment with the outcomes presented in this paper's results section.

When reporting results of a study the confidence interval of a specific confidence level should be reported as well. Confidence interval, or margin of error, is "a term used in inferential statistics that measures the probability that a population parameter will fall between two set values" (Investopedia, 2015, para 1). The confidence interval is reported at a confidence level, which is the probability that the results will occur. Therefore a 95% confidence level with a confidence interval of three would be read as (example) 95% of the population as a whole will fall within a range identified by a plus or minus margin of error of three.

For the second part of this study, which sought to analyze the population identified as students of color it was important to know that the population under study was large enough to produce results, which represented the population as a whole. The complete population studied had 7,077 students of which 1,142 were identified as students of color using the institutions definition of students of color. Of the 1,142 students of color identified, 590 completed to graduation. To state that there is a 95% probability that the study will represent the population as a whole with a confidence interval (margin of error) of +/- 3% there needed to be at least 927 students in the sample who were identified as students of color and of that population at least 552 needed to have graduated (Creative Research Systems, n.d.). The studied population exceeded both requirements therefore this data set can be used to build a predictive model which will produce results that have a 95% confidence level with a confidence interval of three, or in other words there are enough students of color to create a model which should be accurate (+/- 3%) when applied to incoming students of color at the same institution.

Chapter Four

Results

Data Analysis First Year Students

The dataset included any first year student who enrolled during or after Fall 2004 quarter through Summer 2009 quarter and included graduation records through Spring 2015 giving all students at least six years to graduate. This encompasses 7,077 students (n = 7,077). Of those who reported their gender 50.8% were female and 49.2% were male or unreported. The average age at the original admit term was 18.87 years with an age range from 15.5 to 54.92 years old. Race and ethnicity as self-reported were: 62.4% European/Middle Eastern/White; 21.5% not reported; 6.9% Latino/Hispanic; 4.6% Multiracial; 2.2% Asian; 1.8% African American/Black; and less than one percent Alaskan/Native American or Hawaiian/Pacific Islander. Of these 7,077 students studied, 3,845 or 54.33% completed to graduation during this time period. The model target was set as the binomial field that answered (yes, no) did the student graduate.

Decision tree model. After undergoing the iterative process as outlined in the methods section, the decision tree model identified the following student characteristics to be used as variables in the decision tree:

- Developmental math
- First generation
- First quarter financial aid disbursed amount
- High school GPA
- Max GPA from an external institution

- Female
- European/Middle/Eastern/White ethnicity
- Latino/Hispanic ethnicity

The model output produced both a tree graphic (See Figure 3) and the tree nodes as rules (See Appendix C). As identified in the methods section, this model was built using 70% of the data set for training. To evaluate the model it was applied first to the other 30% of the data set for testing the fit of the model and then to the entire data set. The results are shown in Table 3. The testing set had an overall error of 34.76% while the entire data set had an overall error of 35.88%. This indicates that the model is not over-fit to the data and as such is appropriate for use.

Table 3

Decision Tree Model Comparison of Testing Dataset Against Complete Dataset

Model Testing results (on 30% of data) Overall error = 37.46%		
Counts	Predicted	
Actual	No	Yes
No	466	503
Yes	292	861

Model results on complete dataset Overall error = 35.88%		
Counts	Predicted	
Actual	No	Yes
No	1,595	1,637
Yes	926	2,919

Pct.	Predicted	
Actual	No	Yes
No	22 %	24 %
Yes	14 %	41 %

Pct.	Predicted	
Actual	No	Yes
No	23 %	23 %
Yes	13 %	41 %

When analyzing the model for accuracy the testing data set attains a sensitivity of 74.67% with a 95% confidence interval (CI) of 72.06% - 77.16%. When the model is applied to the whole population the sensitivity is 75.92% (95% CI 74.53% - 77.26%).

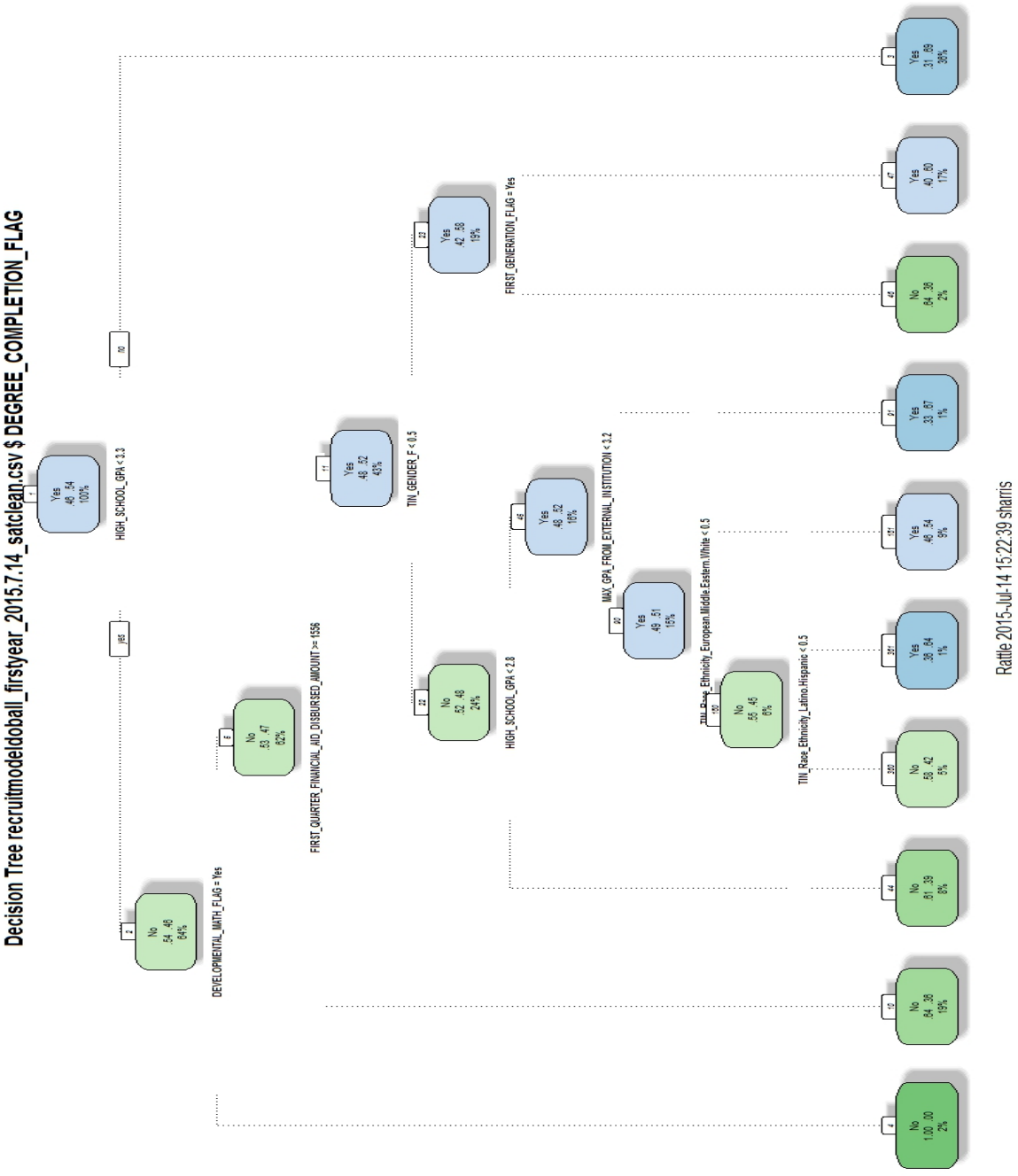


Figure 3: Decision Tree Output for First Year Analysis

Representing the true positive rate, this indicates that 76% of the students who the model predicts will graduate actually do graduate.

The specificity of the testing data set is 48.09% (95% CI 44.90% - 51.29%), while the specificity of the entire data set is 49.35% (95% CI 47.61% to 51.09%). As the true negative rate, this indicates that for those students who the model says will not graduate, it is incorrect about 50% of the time.

Logistic regression model. The logistic regression model was developed with an iterative process of identifying significant characteristics, checking for model error, removing or adding a characteristic until all characteristics identified as significant or the model overall error decreased. Those characteristics (variables) identified as significant were:

- First generation
- High school GPA
- Max GPA from external institution
- Transfer credits accepted
- Transfer test other credits accepted
- Developmental math *
- First quarter financial aid
- Asian *
- European middle eastern white
- Latino/Hispanic
- Female
- First year
- First year transfer

Two characteristics identified in the modeling process were shown not to be statistically significant in the logistic regression however, when the variable(s) were removed and the

regression run, the model error increased. Those variables were * developmental math and self-reported race of * Asian. This remained true whether the model was run with only developmental math, only Asian, or without either. The overall error was lowest when both variables remained in the model therefore the researcher left them in the model. The final logistic regression model has an overall error of 34.59% on the testing data set and an overall error of 35.88% when applied to the entire data set. This small variance indicates that the model has not been over fit. The results are shown in Table 4.

Table 4

Logistic Regression Model Comparison of Testing Data Set Against Complete Data Set.

Model Testing results (on 30% of data) Overall error = 34.59% (N=2122)		
Counts	Predicted	
Actual	No	Yes
No	475	494
Yes	240	913

Results when model applied to complete Overall error = 35.88% (N=7077)		
Counts	Predicted	
Actual	No	Yes
No	1,540	1,692
Yes	847	2,998

Predicted %		
Pct.	Predicted	
Actual	No	Yes
No	22 %	23 %
Yes	11 %	43 %

Predicted %		
Pct.	Predicted	
Actual	No	Yes
No	22 %	24 %
Yes	12 %	42 %

When evaluating the fit of the model, the logistic regression model applied on the testing data set attains a sensitivity of 79.18% (95% CI 76.72% to 81.49%) and when applied to the entire dataset attains a sensitivity of 77.97% (95% CI 76.63% to 79.27%). Again, this indicates that 78% of those students who the model predicts will graduate, actually do graduate.

The specificity of the logistic regression model when applied to the testing dataset is 49.02% (95% CI 45.83% to 52.22%) and 47.65% (95% CI 45.91% to 49.39%) on the entire data set. This specificity indicates that slightly more than half of the students the model says will not graduate, will graduate.

The logistic regression model assigns a probability to each student to indicate their likelihood of graduating. For this study, this ranged from a 2.00931802 to -18.992368 where the higher the number the more likely the probability of the student graduating. To assign a simple prediction of yes/no, zero was used as the cut score with those students who were assigned a probability score equal to or greater than zero predicted to graduate and those students assigned a probability score of less than zero predicted not to graduate.

The logistic regression model predicted 4,690 students (with an assigned probability of zero to 2.009318) would graduate. Of those 4,690 students predicted to graduate, 2,998 actually did graduate (63.92%) and 1,692 did not graduate (36.08%) (See Table 5 and Figure 3). To confirm that zero was the correct cut point to classify students the researcher graphed the logistic regression probability score range from -.15 to the maximum assigned probability of two (See Figure 4).

Table 5

Logistic Regression Model Students Predicted to Graduate Who Did Graduate.

Probability assigned greater than or equal to	Predicted to graduate by model	Did graduate	Logistic regression model accuracy by percentage
-.15	5,218	3,254	62.36 %
-.05	4,894	3,108	63.51 %
0	4,690	2,998	63.92 %
.5	2,440	1,722	70.57 %
1	675	535	79.26 %
1.5	69	61	88.41 %
2	1	1	100 %

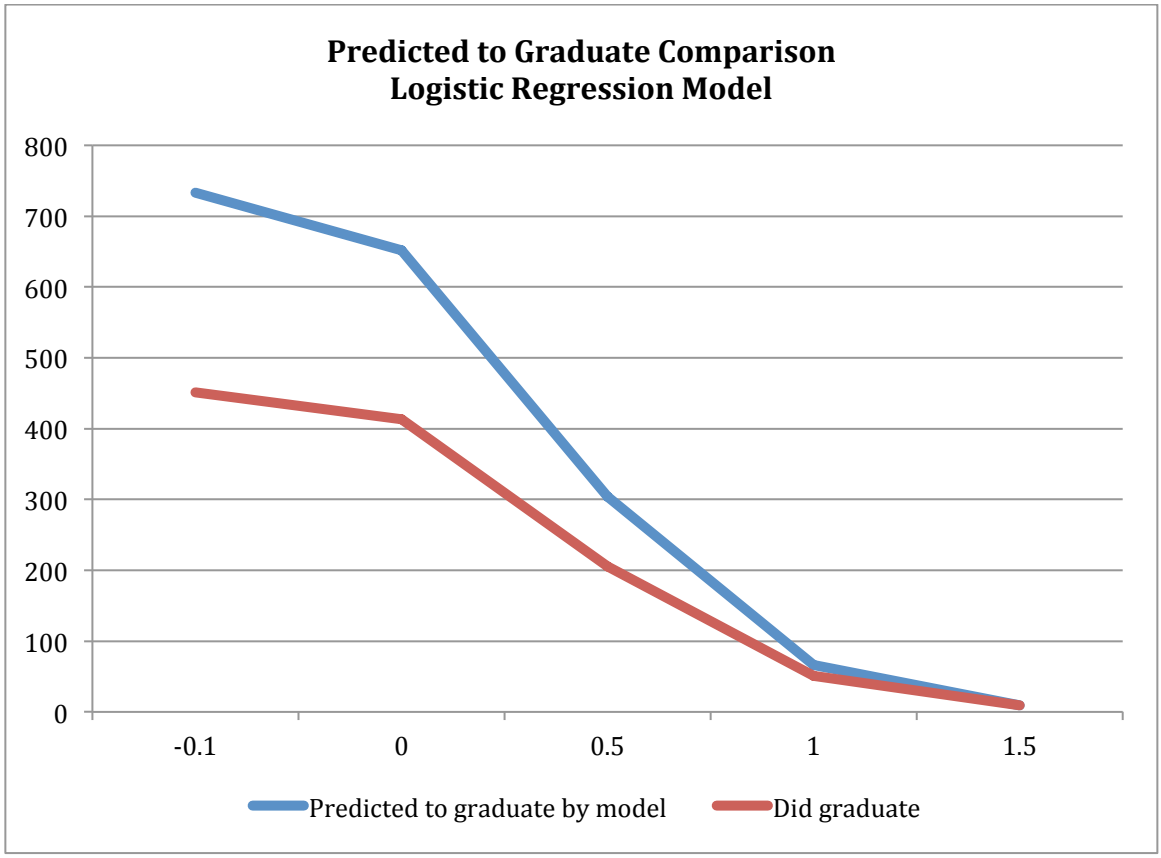


Figure 4: Logistic Regression Comparison – Predicted to Graduate

The logistic regression predicted 2,387 students would not graduate with an assigned probability score of less than zero to -18.992368. Of those 2,387 students, 1,540 did not graduate. Therefore 64.52% were correctly predicted not to complete to graduation. Eight hundred and forty seven (35.48%) of those predicted not to graduate with a probability score of less than zero did complete to graduation. A range of probabilities was analyzed to confirm that zero was the correct cut score to use to assign the prediction yes/no (See Table 6) and the logistic regression probability score range graphed from -.15 to the maximum assigned probability of two (See Figure 4).

Table 6

Logistic Regression Model Comparison, Predicted Not to Graduate

Probability assigned less than zero	Predicted not to graduate by model	Did not graduate	Logistic regression model accuracy by percentage
0	2,387	1,540	64.52 %
-.15	1,859	1,268	68.20 %
-.25	1,583	1,101	69.55 %
-.5	966	721	74.64 %
-1.0	435	355	81.61 %
-1.5	249	217	87.15 %
-2.0	177	171	96.61 %
-3.0	144	144	100 %

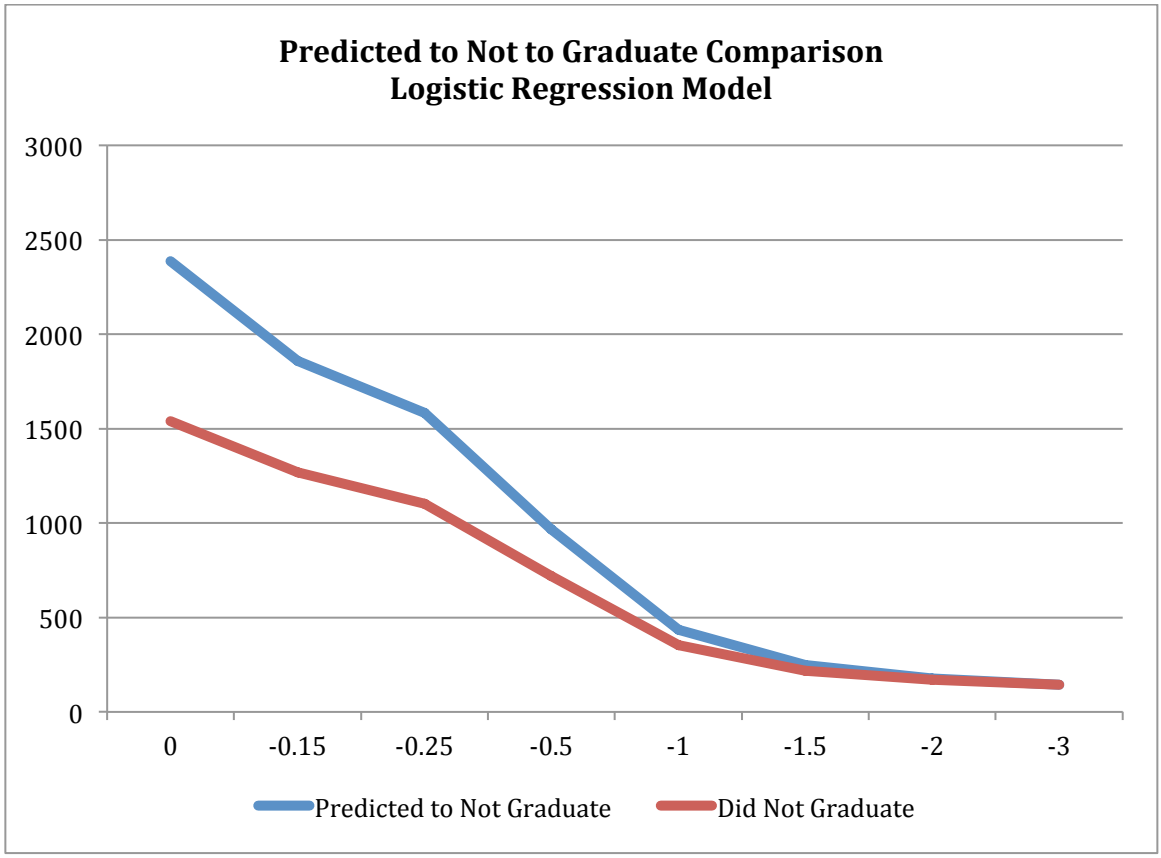


Figure 5: Logistic Regression Comparison – Predicted Not to Graduate

Ensemble models. Using the two predictive models in tandem strengthens the ability of the Division of Enrollment Management (DEM) to identify those students who are most likely to be successful. In tandem, when combining the students that both the decision tree and the logistic regression (>=0) predicted would graduate, the models predict that 3,997 will graduate and 2,658 did graduate, proving the model with an overall accuracy rating of 66.5% (See Table 7).

In tandem, the models sensitivity or true positive rate is 81.93% (95% CI 80.56% to 83.24%). This indicates that in combination and where they agree, 82% of the students

who the two models predict will graduate, do graduate. The specificity of the two models is 48.12% (95% CI 46.18% to 50.07%). Which means that almost half of the time the ensemble models predict the students will not graduate, they do not.

Table 7

Ensemble Model Comparison of Testing Dataset Against Complete Dataset.

Results where models agreed on prediction applied to complete data set Overall error = 33.5%		
Counts	Predicted	
Actual	No	Yes
No	1,242	1,339
Yes	586	2,657
Model agreement	1,828	3,997

Pct.	Predicted	
Actual	No	Yes
No	18 %	19 %
Yes	8 %	38 %

When compared against each other, the models do not agree on 18% of the predicted outcomes (n=1,252). Where the models do not agree the logistic regression model predicted graduation correctly for 49% of the students (340 of 693) while the decision tree model correctly predicted for 47% of the students (261 of 559). Therefore the logistic regression model is slightly more accurate than the decision tree model in the cases where the models do not agree.

Data Analysis First Year Students of Color

The institution under study, in accordance with their strategic plan, seeks to increase the diversity and inclusiveness in their student body. To support this enterprise

initiative, the researcher chose to use the same analysis and predictive modeling techniques on a population identified as students of color using the institution's definition. This dataset included any first year student who enrolled during or after Fall 2004 quarter through Summer 2009 quarter and included graduation records through Spring 2015 giving all students at least six years to graduate. There are 1,142 students who have self-reported as students of color. Of these 576 or 50.4% indicated that they were female and 566 or 49.6% as male or unreported with an average age of 18.77 years upon admit (range 16.83 to 30.46 years). Race and ethnicity as self-reported are shown in Table 8. For the purpose of this research race and ethnicity were analyzed as separate, stand-alone variables; therefore, it is correct that a student of color may have an ethnicity of not Latino/Hispanic with any combination of race other than European/Middle Eastern/White or not reported. It is also correct that a student of color may be European/Middle Eastern/White or not reported as race and be of Latino/Hispanic ethnicity. Of the 1,142 students studied, 590 or 51.67% completed to graduation during this time period. The model target was set as the binomial field DEGREE_COMPLETION_FLAG which answered (yes, no) did the student graduate.

Table 8

Distribution by Ethnicity and Race of Students of Color Population

Race	Ethnicity = Latino/Hispanic	Ethnicity = Not Latino/Hispanic	Ethnicity = Not reported
African American/Black	2 (.2%)	104 (9.1%)	23 (2.0%)
Alaska/Native American	2 (.2%)	28 (2.5%)	8 (0.7 %)
Asian	0	135 (11.8%)	22 (1.9 %)
European/Middle Eastern/White	9 (0.7%)	0	0
Hawaiian/Pacific Islander	1 (0.1%)	15 (1.3%)	0
Multiracial	91 (8.0%)	300 (26.3%)	23 (2.0%)
Not reported	379 (33.2%)	0	0

Decision tree model. After undergoing the process described in the methods section, the student of color decision tree model identified the following student characteristics (variables) to be used in the model:

- Developmental math
- First generation
- First quarter financial aid disbursed amount
- Highest SATACT composite
- High school GPA
- Latino Hispanic ethnicity
- Not Latino/Hispanic ethnicity
- Admit type FYR
- Transfer credits accepted

Figure 5 shows the model output tree graphic as tree nodes while the tree as rules was used to develop the formulas (See Appendix C).

Seventy percent of the data set was used for model training while 30% quarantined for testing. Table 9 shows the results of the testing data set compared to the results when the model was applied to the entire student of color dataset, n = 1,142.

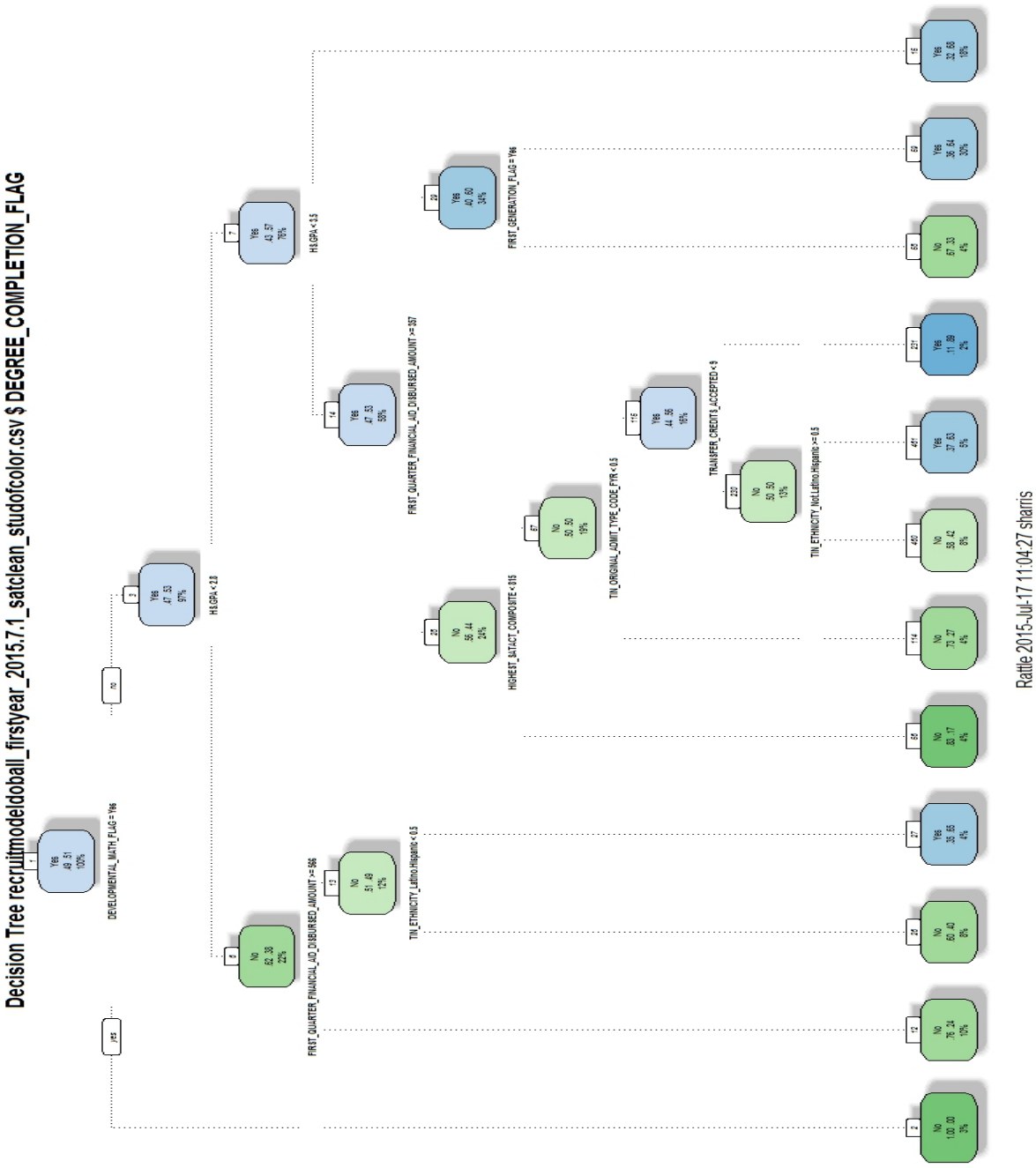


Figure 6: Decision Tree Output for First Year Student of Color Analysis

Table 9

Student of Color Decision Tree Model Comparison of Testing Dataset Against Complete Dataset

Model Testing results (on 30% of data) Overall error = 37.46%		
Counts	Predicted	
Actual	No	Yes
No	88	72
Yes	46	136

Results when model applied to complete Overall error = 35.88%		
Counts	Predicted	
Actual	No	Yes
No	317	235
Yes	140	450

Pct.	Predicted	
Actual	No	Yes
No	26 %	21 %
Yes	13 %	40 %

Pct.	Predicted	
Actual	No	Yes
No	28 %	21 %
Yes	12 %	39 %

Evaluating this model for validity shows the testing dataset attains a sensitivity of 74.73% (95% CI 67.76 – 80.86%). When applied to the student of color population the sensitivity attains a slightly better rate of 76.27% (95% CI 72.63% - 79.65%). This indicates that of those students who the model predicts to graduate, 76% will graduate.

The specificity of the testing dataset returns a value of 55.00% (95% CI 46.95% - 62.86%). The entire student of color dataset attains a specificity of 57.43% (95% CI 53.18% - 61.59%). The specificity indicates that of the students the model predicts will not graduate, 57% do not graduate.

Logistic regression model. Using the iterative process outlined in the methods section, the following characteristics were identified as statistically significant:

- High school GPA
- Transfer credits accepted
- Max GPA from external institution
- First quarter financial aid disbursed amount
- First generation flag – yes
- Highest SATACT composite
- Developmental math *
- Latino/Hispanic ethnicity
- Not Latino/Hispanic ethnicity
- Female
- First year

As was true with the original dataset, developmental math was not shown to be statistically significant by the logistic regression, however when it was removed the overall standard error of the model increased. Due to that and the relevance developmental math presented in the decision tree model the researcher left it as a variable in this logistic regression model. The testing set returned an overall error of 34.59% while the entire student of color dataset had a standard error of 35.88% (See Table 10). This small variance indicates the model has not overlearned and is a good fit. As was explained in the methods section, actual evaluation of the model is more directly served using sensitivity and specificity.

Evaluation of the logistic regression model on the student of color testing dataset shows sensitivity (the true positive rate) attains 67.03% (95% CI 59.69% - 73.81%) and when applied to the entire dataset improves to 70.00% (95% CI 66.12% - 73.67%). This indicates that 70% of the students the model predicts will graduate, do graduate.

Table 10

Student of Color Logistic Regression Model Comparison of Testing Dataset Against Complete Dataset

Model Testing results (on 30% of data) Overall error = 34.59%		
Counts	Predicted	
Actual	No	Yes
No	90	70
Yes	60	122

Results when model applied to complete Overall error = 35.88%		
Counts	Predicted	
Actual	No	Yes
No	313	239
Yes	177	413

Pct.	Predicted	
Actual	No	Yes
No	26 %	20 %
Yes	18 %	36 %

Pct.	Predicted	
Actual	No	Yes
No	27 %	21 %
Yes	16 %	36 %

The specificity of the logistic regression model when applied to the testing dataset is 56.25% (95% CI 48.20% - 64.07%) and 56.70% (95% CI 52.45% - 60.88%) when applied to the entire dataset. This indicates that over half of the students who are predicted not to graduate, will not graduate.

The logistic regression assigns a probability index to each student ranging from 1.947527 to -17.5032. The higher the number the more likely the student will graduate. For the purpose of assigning a yes/no predictor, a cut score of zero was used. Those students with a logistic regression probability equal to or greater than zero are predicted to graduate while those with a probability score less than zero are predicted not to graduate. The model predicted that 652 would graduate, of which 413 (63.34%) did graduate while 239 (36.66%) did not graduate. The researcher graphed a logistic

regression probability score range from -.1 to 1.5 to verify that zero was a correct cut score (See Table 11 and Figure 6).

Table 11

Logistic Regression Model Students of Color Predicted to Graduate Who Did Graduate

Probability assigned greater than or equal to	Predicted to graduate by model	Did graduate	Logistic regression model accuracy by percentage
-.1	733	451	61.53 %
0	652	413	63.34 %
.5	304	205	67.43 %
1	66	51	77.27 %
1.5	9	9	100 %

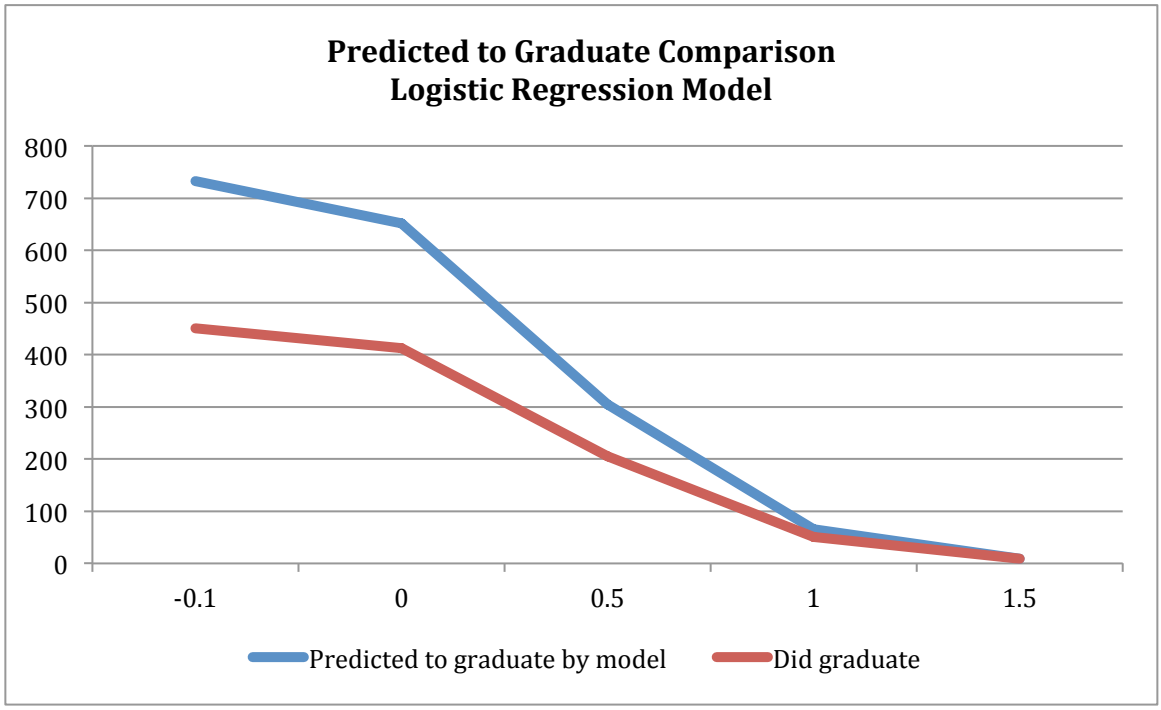


Figure 7: Logistic Regression Comparison – Predicted to Graduate

Table 12

Logistic Regression Model Students of Color Predicted Not to Graduate Who Did Not Graduate

Probability assigned less than zero	Predicted not to graduate by model	Did not graduate	Logistic regression model accuracy by percentage
0	490	313	63.27
-.15	377	256	67.91
-.25	316	221	69.94
-.5	204	149	73.04
-1.0	82	65	79.27
-1.5	43	40	93.02
-2.0	33	32	96.97
-3.0	28	27	96.43

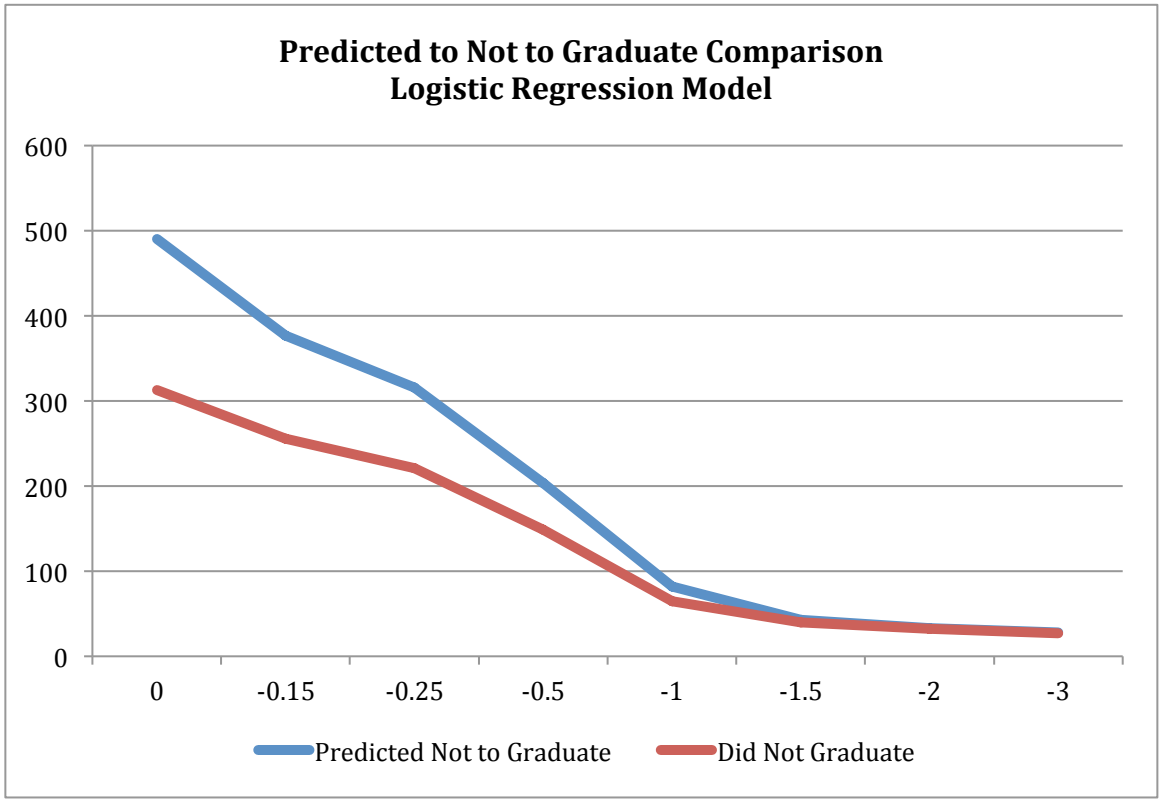


Figure 8: Students of Color Logistic Regression Comparison – Predicted Not to Graduate

Ensemble models: Combining the two models into an ensemble model, basically using them in tandem, strengthens the ability of the DEM to target those students who are most likely to be successful. For the first year students of color dataset where combined, the ensemble model predicted that 540 of the 1,142 students would graduate. Three hundred and sixty six did graduate giving the ensemble model an overall accuracy of 67.8 % (See Table 13).

The ensemble model attains a sensitivity or true positive rate of 79.91% (95% CI 75.95% - 83.49%). This indicates that where the models are in agreement, 80% of the students that they predict will graduate, do graduate. The specificity of the ensemble model is 59.25% (95% CI 54.42% - 63.95%). This means almost 60% of the time the model is correct when it predicts a student will not graduate (true negative rate).

Table 13

Student of Color Ensemble Model Comparison of Testing Dataset Against Complete Dataset

Results when model applied to complete Overall error = 33.5%		
Counts	Predicted	
Actual	No	Yes
No	253	174
Yes	92	366
Model agreement	345	540

Predicted %		
Actual	No	Yes
No	29 %	20 %
Yes	10 %	41 %

When the two models are compared, 22.5% of the time (257 students) they do not agree as to the predicted outcome. They disagree on the outcome of 112 students in the logistic regression model and 145 students in the decision tree model. Where they disagree, the decision tree model is correct 59% of the time (85 of 145 students graduate) whereas the logistic model is correct only 43% of the time (48 of 112 students graduate). Therefore when not in accordance, the more accurate model is the decision tree model.

Results Summary

The decision tree is a binary predictor of student graduation. For the entire population, the model returned an overall error of 35.9%, attained a sensitivity rate of 75.9% and a specificity rate of 49.4%. The student of color population model returned an overall error of 35.9% with a sensitivity rate of 76.3% and a specificity rate of 57.4%.

The logistic regression resulted a probability score for each student, the higher the probability score, the more likely that student will earn a degree. The overall error was 35.9% on the entire dataset, with a sensitivity rate of 78.0%, and a specificity rate of 47.7%. The student of color population model returned an overall error of 35.9% with a sensitivity rate of 70.0% and a specificity rate of 56.7%.

When combining the two models for the entire population there was disagreement on only 18% of the predicted outcomes and an overall error of 33.5%. More importantly, when the models agree that a given student is predicted to graduate, 82% of those predicted to complete, earned a degree. When the models agree that the student will not

graduate, 48.1% do not graduate. Upon review of the students of color model, there is disagreement on only 23% of the predicted outcomes with an overall error of 32.2%. As with the larger model, when in agreement the models correctly predict those students who graduate 80% of the time. Interestingly the students of color model returns a higher specificity rate than the larger population model by predicting those students who will not graduate correctly almost 60% of the time.

Table 14

Ensemble Model Evaluation

Ensemble Predictive Models – Whole Population				
Results when model applied to complete Overall error = 33.5% (N=7,077)				Sensitivity (True Positive Rate)
	Predicted			81.9%
Actual	No	Yes		Specificity (True Negative Rate)
No	1,242 (18%)	1,339 (19%)		
Yes	586 (8%)	2,657 (38%)		48.1%
Agreement	1,828 (26%)	3,996 (57%)		Not in Agreement 18%
Ensemble Predictive Models – Students of Color Population				
Results when model applied to complete Overall error = 32.2% (N=1,142)				Sensitivity (True Positive Rate)
	Predicted			79.9%
Actual	No	Yes		Specificity (True Negative Rate)
No	253 (22%)	174 (15%)		
Yes	92 (8%)	366 (32%)		59.3%
Agreement	345 (30%)	540 (47%)		Not in Agreement 23%

Chapter Five

Discussion and Recommendations

The results of this case study are robust with implications for use in the suggested avenues of intentional recruiting and tuition yield predictions; as well as retention and graduation predictions. Additionally administration can use the models to develop potential intervention resource allocation plans for those students identified as having a lower probability of completing to earn a degree. As hypothesized by the researcher, specific characteristics found to be statistically significant have been identified to describe those students who enroll, persist, and graduate from the institution under study. Using these identified characteristics it is possible to predict any given incoming first year undergraduate student's probability of completing through degree attainment and graduation with an overall accuracy greater than 66%. Grouping, or clustering these characteristics into profiles facilitates recruiting activities, whereby those students more likely to succeed can be intentionally sought and engaged regarding attending this institution.

The idea of profiling any group has a negative connotation in our current socio-political atmosphere. The use of the term student profile here should not be tainted with negativity, as nowhere does the researcher suggest that any individual or group of students be ignored, discouraged, discriminated against nor repudiated. In fact, the researcher recommends against changing any existing business processes currently in

place specifically to ensure that those students who currently enroll at, or are considering enrolling at, the institution under question and who would not be identified by the predictive models as a potentially successful candidate are not disenfranchised. For the purpose of this study, the term student profile is used simply to describe a list of attributes that allows clustering of that student into a group of similar like students.

Potential Uses of Study Results

Creating student profiles for the purposes of classifying potential and incoming students probability of graduation enable distinct and separate potential business processes. Specifically, the researcher recommends four uses for the results of this study.

Intentional recruiting. The researcher hypothesized that targeted recruiting efforts based on these recommendations will recognize a substantial ROI in the form of higher graduation rates, focused recruiting expenditures, and improved institution awareness and reputation by constituents. The DEM may use the student profiles created to identify those, applicant and non-applicant alike, potential students who are more likely to be successful and reach out to them to purposefully recruit them.

As outlined in the methods section, sensitivity is used to evaluate the model to find the extent to which the model accurately predicts what it is designed to predict. The sensitivity for the predictive models built was quite strong at 80-82%. What this means is that 80% of the students of color and 82% of the population as a whole that the models predict will persist and complete to degree – do.

In a 2012 report to the Office of Financial Management of the State of Washington, this institution reported a six-year graduation rate for first year students of 55.8% (OFM, p. 3). Fall 2015 first year student enrollment was reported at 1,653, of whom 922 or 55.8% will graduate in six years if graduation rates remain steady (Central Washington University, 2015). If only 100 additional students were recruited using the characteristics identified by this study, next year's incoming class would consist of 1,753 new students of whom 1,004 would graduate (922 at 55.8% + 82 at 82%) increasing the overall six-year graduation rate by 1.5% to 57.3%. These additional 26 graduating students who denote the 82 students who graduate based on the intentional recruiting model less the 56 students who would graduate based on current methods represent increased net new tuition funding of $163,800 annually or $982,800 over the six-year course of study.

While sensitivity measures the true positive rate, specificity is the probability that the model correctly predicts the true negative rate. The specificity rates returned by the models were lower than desired; however, this is not surprising as the study was designed to identify those students who would complete rather than those who would not. The specificity rates, 48% for all first year students and 59% for first year students of color, indicate that it is important for the DEM to combine intentional recruiting efforts with current business processes rather than replacing current practices. Should the DEM choose to replace existing efforts with only targeted recruiting efforts there is a high-risk that persistence and graduation rates would actually drop. This is due to the high

specificity rates, which indicate approximately 50% of the students that the study predicts, will not graduate, actually do graduate.

Identify high-risk students. The Student Success Division may use the predictive models to identify those at-risk students who have lower probability scores (and therefore a higher risk of non-completion). Student outreach interventions should be developed and operationalized providing additional resources to these at risk students. Assuming these at-risk students are reached prior to the critical point where the student has chosen to stop/drop out, retention and persistence rates will increase thereby increasing graduation rates. The effectiveness of these interventions will naturally necessitate evaluating the predictive models annually for fit and updating them as needed.

Use predictive model to plan needed resources. The Student Success Division and administration may use the predictive models to plan additional resources needed to support recommended intervention and at risk student resource programs. These programs can be budgeted for based on the percentage of incoming students identified by the predictive model as having a lower probability of completing successfully score.

Predicting tuition yield. The Division of Enrollment Management may use the predictive models to report to administration projected tuition yields built on the probability scores assigned to a student based on the profile the student fits. Each fall, the appropriate models can be applied to the incoming first year cohort, producing projected enrollment yields which can be converted to projected tuition yield and provided to Business and Financial Affairs as well as upper administration. As additional

support and intervention resources are placed in service it will be important to assess the models fit and update the models annually.

A higher degree of confidence should be recognized using the predictive models to develop projected tuition yield. The research shows that 80% of those positively indicated students (sensitivity - yes/yes) will graduate, and therefore be paying tuition. The research also shows that about 50% of the students that the models predict will not graduate (specificity - no/no) will also graduate, and therefore will also be paying tuition. As such these projected enrollments can be used to create formulas and develop tuition revenue models, an effort which is beyond the scope of this research.

Implementation Concerns

The researcher recognizes operationalizing the findings of this study will require the institution at hand to implement an iterative process as outlined in Figure 8. Failure to adopt all steps of the cycle will limit the effectiveness of the model, especially over time. Without increasing resources and creating outreach to at-risk populations, persistence and graduation rates will increase only marginally. Failure to evaluate and update the model annually render the projections inaccurate and create the opportunity to adversely affect persistence and graduation rates through the use of outdated student characteristics and resulting probabilities.

Future Studies

Potential futures studies for consideration should include an update and analysis of the effectiveness of the predictive models developed after implementation of intentional recruiting strategies and/or at-risk student intervention programs. Other possibilities include researching potential additional attributes for analysis to be included in the predictive models which may include such things as social media interactions and relationships based on institution alumni, proximity to institution, or program of interest.

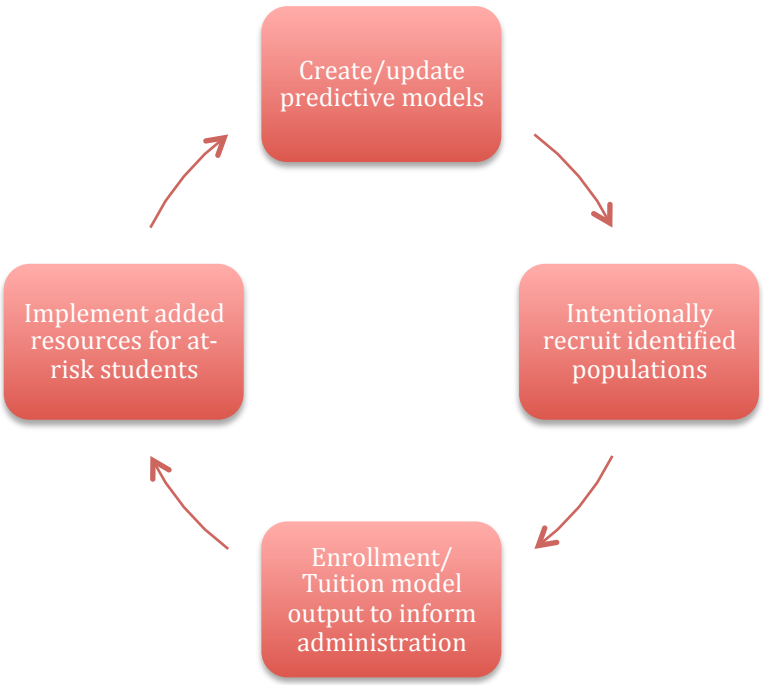


Figure 9: Lifecycle of Operationalizing the Study Findings

Chapter Six

Conclusion

Decreased state and federal funding of post-secondary public institutions in the United States, over the last ten to fifteen years, has transformed their business model to one where they must now rely heavily on tuition revenue as a major source of income. Improvements in recruiting, predictive enrollment modeling, and accuracy in tuition yield projections advance administrative functions and provide needed information to student support services. The researcher sought to provide actionable information in support of these initiatives by using BI techniques to answer four research questions:

1. Can statistically significant characteristics be identified to provide a basis for intentional recruiting at the institution under study?
2. Using these characteristics, can successful completion of an undergraduate degree by a first year student be predicted?
3. What are the characteristics identified?
4. Are the characteristics different in populations that identify as diverse by race and/or ethnicity?

Research Questions Answered

1. Can statistically significant characteristics be identified to provide a basis for intentional recruiting at the institution under study? Yes, using analytics the

researcher identified those statistically significant characteristics generally known about potential applicants prior to the recruiting process. The decision trees analysis provided a list of eight characteristics while the logistic regression analysis identified 13 attributes. The models, when used in tandem, strengthen the ability of the recruiters to identify those students who are most likely to be successful.

2. Using these characteristics, can successful completion of an undergraduate degree by a first year student be predicted? Yes, undertaking an iterative process the researcher created models to predict first year undergraduate degree attainment within six years. The models were created using decision tree analysis and logistic regression and finally an ensemble predictive model combined the two previously developed models. The sensitivity rate indicates that if students with identified characteristics are intentionally recruited, 80-82% of them will be successful and earn a degree within six years.

3. What are the characteristics identified? The characteristics found to be statistically significant in predicting enrollment and completion and that were used to build the logistic regression and decision tree predictive models are shown in Table 15.

4. Are the characteristics different in populations that identify as diverse by race and/or ethnicity? Yes, the significant characteristics for the student of color models differed from the population as a whole. The characteristics found to be statistically significant in predicting degree attainment were used to build the decision tree and logistic regression predictive models. These characteristics are shown in Table 16.

Table 15

Significant Characteristics Identified to Predict First Year Undergraduate Degree Attainment at the Case Study Institution

Decision Tree Model	Logistic Regression Model
Developmental math First generation First quarter financial aid disbursed amount High school GPA Max GPA from an external institution Female European middle eastern white ethnicity Latino Hispanic ethnicity	First generation High school GPA Max GPA from external institution Transfer credits accepted Transfer test other credits accepted Developmental math First quarter financial aid Asian European middle eastern white Latino/Hispanic Female First year First year transfer

Table 16

Significant Characteristics Identified to Predict First Year Students of Color Undergraduate Degree Attainment at the Case Study Institution

Decision Tree Model	Logistic Regression Model
Developmental math First generation First quarter financial aid disbursed amount Highest SATACT composite High school GPA Latino Hispanic ethnicity Not Latino/Hispanic Admit type FYR Transfer credits accepted	High school GPA Transfer credits accepted Max GPA from external institution First quarter financial aid disbursed amount First generation flag – yes Highest SATACT composite Developmental math Latino/Hispanic Not Latino/Hispanic Female First year

Actionable Information

The Division of Enrollment Management at the institution under study has two main responsibilities: to recruit new students and to report enrollment to administration. The models built in this study can be used to support both obligations. Using the student characteristic profiles identified to intentionally recruit new student populations will help to meet enrollment targets. In addition, because the models have been built based on students who enroll and complete to degree, retention, persistence and completion rates will increase. Tuition and enrollment yields will be more accurately reported for a broader time period. Applying the predictive models to incoming first year student cohorts will provide the DEM with the probability of each student completing a degree within the next six years, which can be used to create both enrollment yield and tuition yield formulas.

The various student support functions on campus will be able to plan and budget for needed resources based on the probabilities assigned to incoming students by the predictive models. At risk scores can be determined, the size of the population extrapolated and additional resources planned for those students who require additional ancillary services to be successful.

Synopsis

In conclusion, this study has achieved the research goal of identifying statistically significant characteristics to be used when recruiting first year students at a division II, public comprehensive, university in Washington State. The predictive models built accurately predict 80-82% of those first year undergraduate students who complete to degree within six years at the institution under study. The researcher has outlined how the predictive models can be used to improve enrollment yield and tuition yield projections as well as to plan and budget for additional student support resources. Potential shortcomings, and future studies have been noted.

References

1Keydata.com. (n.d.). Data Warehouse Definition – What is a Data Warehouse? Retrieved from http://www.1keydata.com/datawarehousing/data-warehouse-definition.html

ACT. (2013). ACT – SAT Concordance: A Tool for Comparing Scores. Retrieved from http://www.act.org/aap/concordance/pdf/reference.pdf

Adelman, C. (2006). The Toolbox Revisited Paths to Degree Completion from High School Through College. Retrieved from http://files.eric.ed.gov/fulltext/ED490195.pdf

Amburgey, W., & Yi, J. (2011). Using Business Intelligence in College Admissions: A Strategic Approach. *International Journal of Business Intelligence Research, 2(1),* 1-15 doi: 10.4018/jbir.2011010101

Antons, C., & Maltz, E. (2006). Expanding the Role of Institutional Research at Small Private Universities: A Case Study in Enrollment Management Using Data Mining. *Data Mining in Action Case Studies in Enrollment Management, 131,* 69-81. doi: 10.1002/ir.188

Asllani, A. (2014, November). Business Analytics with Management Science Models and Methods (FT Press Analytics). Pearson Education. Kindle Edition.

Bichsel, J. (2012, June 22). Analytics in Higher Education: Benefits, Barriers, Progress, and Recommendations. *Educause Research Hub.* Retrieved from http://www.educause.edu/library/resources/2012-ecar-study-analytics-higher-education

Brown, C., Dehayes, D., Hoffer, J., Martin, E., & Perkins, W. (2012). *Managing Information Technology*. Boston, MD: Prentice Hall.

BusinessDictionary.com. (n.d.). *What is data analytics? Definition and meaning*. Retrieved from http://www.businessdictionary.com/definition/analytics.html

BusinessDictionary.com. (n.d.). *What is data warehouse? Definition and meaning*. Retrieved from http://www.businessdictionary.com/definition/data-warehouse.html

Central Washington University. (n.d.). *Student Success.* Retrieved from http://www.cwu.edu/student-success/

Central Washington University. (2015, October 27). *CWU Freshmen Enrollment Continues Dramatic Upward Trend.* Retrieved from http://www.cwu.edu/cwu-freshmen-enrollment-continues-dramatic-upward-trend

Cerna, O., Perez, P., & Saenz, V. (2009, April). Examining the Precollege Attributes and Values of Latina/o Bachelor's Degree Attainers. *Journal of Hispanic Higher Education, 8(2),* 130-157. doi: 10.1177/1538192708330239

Chang, L. (2006). Applying Data Mining to Predict College Admissions Yield: A Case Study. *Data Mining in Action Case Studies in Enrollment Management, 131,* 53-68. doi: 10.1002/ir.187

Chen, C. (2008, January 18). An integrated Enrollment Forecast Model. *IR Applications, 15,* 1-18. Retrieved from http://files.eric.ed.gov/fulltext/ED504328.pdf

Churn-rate.com. (n.d.). Churn Rate 101. Retrieved from http://churn-rate.com

Creative Research Systems. (n.d.). Sample Size Calculator. Retrieved from http://www.surveysystem.com/sscalc.htm#one

DataCookbook. (n.d.). *Data Cookbook – Central Washington University*. Retrieved from DataCookbook Database.

DataCookbook. (n.d.). *Data Cookbook – Common Data Set*. Retrieved from DataCookbook Database.

DataCookbook. (n.d.). *Data Cookbook – Predictive Analytics Reporting Framework*. Retrieved from DataCookbook Database.

DataCookbook. (n.d.). *Data Cookbook – Texas Higher Education Coordinating Board*. Retrieved from DataCookbook Database.

Donalek, C. (2011, April). Supervised and Unsupervised Learning. Retrieved from http://www.astro.caltech.edu/~george/aybi199/Donalek_Classif.pdf

Epstein, J. (2010). The Creation of Enrollment Management at Boston College: A History as told by the original Enrollment Management Team. Retrieved from http://www.maguireassoc.com/wp-content/uploads/2012/03/Creation-of-Enrollment-Management.pdf

Faircloth, S., Alcantar, C., & Stage, F. (2015). Use of Large-Scale Data Sets to Study Educational Pathways of American Indian and Alaska Native Students. *New Scholarship in Critical Quantitative Research Part 2, 163,* 5-24. doi: 10.1002/ir.20083

Gartner. (n.d.). *Business Intelligence.* Gartner IT Glossary. Retrieved from http://www.gartner.com/it-glossary/business-intelligence-bi/

Gartner Group. (2000). *The GartnerGroup CRM Glossary*. Retrieved from http://www.gartnerweb.com/public/static/hotc/hc00086148.html

Goldstein, P. (2005). Academic Analytics: The Uses of Management Information and Technology in Higher Education. *ECAR Key Findings.* Louisville, CO: ECAR, December 2014. Available from http://www.educause.edu/ecar

Government Accountability Office. (2014). Higher Education State Funding Trends and Policies on Affordability. (GAO Publication No. 15-151). Washington, DC.: U.S. Government Printing Office.

Hanover Research. (2011, May). Predicting College Student Retention. Retrieved from http://www.algonquincollege.com/student-success/files/2014/12/Predicting-College-Student-Retention-Literature-Review-1.pdf

Hossler, D. & Bean, J. P., & Associates. (1990). *The Strategic Management of College Enrollments.* San Francisco: Josey-Bass Publishers.

Hutchinson, V., & Quintas, P., (2008). Do SMEs do Knowledge Management? Or Simply Manage what they Know? *International Small Business Journal*, *(26)2,* 131-154.

Inman, E. & Mayes, L. (1999). The Importance of Being First: Unique Characteristics of First Generation Community College Students. *Community College Review, 26(4),* 3-22. doi: 10.1177/009155219902600402.

Investopedia. (2015). Confidence Interval Definition. Retrieved from http://www.investopedia.com/terms/c/confidenceinterval.asp

Janssen, C. (n.d.). *Extract Transform Load – Definition*. Techopedia. Retrieved from http://www.techopedia.com/definition/24170/extract-transform-load-etl

Johnson, I. (2008). Enrollment, Persistence and Graduation of In-State Students at a Public Research University: Does High School Matter?. *Research in Higher Education, 49(8),* 776-793. doi: 10.1007/s11162-008-9105-8.

Justice, K. (2013, March 19). Budget and Policy Center, Washington State Cuts to Higher Education Among Worse in The Country. *Washington State Budget and Policy Center: Schmudget Blog.* Retrieved from http://budgetandpolicy.org/schmudget/washington-state-cuts-to-higher-education-among-worst-in-the-country

Justice, K. (2011, April 26). Undermining Prosperity: Higher Education Cuts Weaken Access, Affordability and Quality. *Washington State Budget and Policy Center.* Retrieved from http://budgetandpolicy.org/policy-areas/state-budget/education

Kurz, K. and Scannell, J. (2006, May). Enrollment Management Grows Up. *University Business*. Retrieved from http://www.universitybusiness.com/article/enrollment-management-grows

Lang, L. & Pirani, J. (2014, April 23). BI Reporting, Data Warehouse Systems and Beyond. *Research bulletin.* Louisville, CO: ECAR. Available from http://www.educause.edu/ecar

Larson, J. (2013, July 12). The history of higher education marketing: 1976 and John Maguire's enrollment management. *U of Admissions Marking: Digital enrollment trends, techniques, and strategies for universities and colleges*. Retrieved from http://www.uofadmissionsmarketing.com/2013/07/the-history-of-higher-education.html

Lotkowski, V., Robbins, S., & Noeth, R. (2004). The Role of Academic and Non-Academic Factors in Improving College Retention. *ACT Policy Report*. Retrieved from https://www.act.org/research/policymakers/pdf/college_retention.pdf

Luan, J. (2002). Data Mining and its Applications in Higher Education. *Knowledge Management Building a Competitive Advantage in Higher Education*. (pp. 17-36). San Francisco, CA: Josey-Bass.

MedCalc.org. (2015, August 14). Diagnostic test evaluation. *MedCalc.* Retrieved from https://www.medcalc.org/calc/diagnostic_test.php

Merriam-Webster. (n.d.). Definition of persistence. Retrieved from http://www.merriam-webster.com/dictionary/persistence

NG Data. (n.d.). What is Predictive Analytics? Definition and Models. Retrieved from http://www.ngdata.com/what-is-predictive-analytics/

Nandeshwar, A., & Chaudhari, S. (2009, April 22). Enrollment Prediction Models Using Data Mining. Retrieved from http://nandeshwar.info/wp-content/uploads/2008/11/DMWVU_Project.pdf

Noel-Levitz. (2013). 2013 Cost of recruiting an undergraduate student: Benchmarks for four-year and two-year institutions. Coralville, Iowa: Authors. Retrieved from http://www.noellevitz.com/BenchmarkReports

Norris, D. & Baer, L. (2013, February 25). Building Organizational Capacity for Analytics. *Educause Publications*. Retrieved from http://www.educause.edu/library/resources/building-organizational-capacity-analytics

Office of Financial Management, OFM. (2012). *Central Washington University.* Retrieved from http://www.ofm.wa.gov/performance/plans/2012_CWU.pdf

Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, *56*(1), 45–50.

Recruit. 2015. In *Dictionary.com Unabridged.* Retrieved from http://dictionary.reference.com/browse/recruit

Rios-Aguilar, C. (2015). Using Big (and Critical) Data to Unmask Inequities in Community Colleges. *New Scholarship in Critical Quantitative Research Part 2, 163,* 43-57. doi: 10.1002/ir.20085.

Serban, A., & Luan, J. (2002). Editors' Notes. *Knowledge Management Building a Competitive Advantage in Higher Education*. (pp. 1-3). San Francisco, CA: Josey-Bass.

Siegel, E. (2013, February). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Wiley. Kindle Edition.

Statistics Solutions. (2015). *Using Logistic Regression in Research.* Retrieved from http://www.statisticssolutions.com/using-logistic-regression-in-research/

Steinberg, J. (2010, May 12). The Early Line on Admission Yields (and Wait-List Offers). *The New York Times.* Retrieved from http://thechoice.blogs.nytimes.com/2010/05/12/yield-3/?_r=0

Tutorials Point. (n.d.). *Data Warehouse Tutorial.* Retrieved from http://tutorialspoint.com/dwh/dwh_pdf_version.htm

University of Strathclyde. (n.d.). What is Logistic Regression?. Retrieved from https://www.strath.ac.uk/aer/materials/5furtherquantitativeresearchdesignandanalysis/unit6/whatislogisticregression/

U.S. Department of Education (DOE), Federal Student Aid Department. (2013, May 16). Time Limitation on Direct Subsidized Loan Eligibility for First-Time Borrowers on or after July 1, 2013. Retrieved from https://ifap.ed.gov/eannouncements/attachments/051613DirectSubsidizedLoanLimit150PercentAnnounce1Attach.pdf

U.S. Department of Education, National Center for Educational Statistics. (n.d.). The Integrated Postsecondary Education Data System – Glossary. Retrieved from http://nces.ed.gov/ipeds/glossary/?text=1

Van Barneveld, A., Arnold, K., and Campbell, J. (2012, January). Analytics in Higher Education: Establishing a Common Language. *Educause Learning Initiative*. Retrieved from http://net.educause.edu/ir/library/pdf/eli3026.pdf

Webopedia. (n.d.). What is Online Analytical Processing OLAP? *Webopedia.* Retrieved from http://www.webopedia.com/TERM/O/OLAP.html

Williams, G., (2010). Data Mining, Desktop Survival Guide. Retrieved from http://datamining.togaware.com/survivor/Complexity_cp.html

Zhao, C. & Luan, J. (2006). Data Mining: Going Beyond Traditional Statistics. *Data Mining in Action Case Studies of Enrollment Management, 131,* 7-16. doi: 10.1002/ir.184.

Appendices

Appendix A: Glossary of Terms

Admitted students: Applicants that have been granted an official offer to enroll in a postsecondary institution (U.S. Department of Education, National Center for Educational Statistics[NCES], n.d.).

Advanced placement (AP) courses: College-level courses taught in high school. Students may take an examination at the completion of the course; acceptable scores allow students to earn college credit toward a degree, certificate, or other formal award (NCES, n.d.).

American Indian or Alaska Native: A person having origins in any of the original peoples of North and South America (including Central America) who maintains cultural identification through tribal affiliation or community attachment (NCES, n.d).

Analytics: "Analytics often involves studying past historical data to research potential trends, to analyze the effects of certain decisions or events, or to evaluate the performance of a given tool or scenario" (BusinessDictionary.com, n.d.).

Applicant: An individual who has fulfilled the institution's requirements to be considered for admission (including payment or waiving of the application fee, if any) and who has been notified of one of the following actions: admission, non-admission, placement on waiting list, or application withdrawn by applicant or institution (NCES, n.d.).

Asian: A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian Subcontinent, including for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam (NCES, n.d.).

Black or African American: A person having origins in any of the black racial groups of Africa (NCES, n.d).

Business Intelligence: "is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance" (Gartner IT, n.d.)

Cohort: A specific group of students established for tracking purposes (NCES, n.d.).

Completer: A Student who receives a degree, diploma, certificate, or other formal award. In order to be considered a completer, the degree/award must actually be conferred (NCES, n.d.).

Completion: Status achieved upon successful completion of all requirements in a degree program. For the purpose of this paper Completion refers specifically to completion of requirements in an approved course of study, which qualifies the student to earn a degree.

Course of study: Any grouping of courses, which are represented as entitling a student to a degree or certificate. Also known as program or program of study (Data Cookbook – Texas Higher Education Coordinating Board, n.d.).

Data warehouse: "Massive database serving as a centralized repository of all data generated by all departments and units of a large organization. Advanced data mining

software is required to extract meaningful information from a data warehouse" (BusinessDictionary.com, n.d.)

Degree: An award conferred by a college, university, or other postsecondary education institution as official recognition for the successful completion of a program of studies (NCES, n.d).

Degree-seeking: Students enrolled in courses for credit who are recognized by the institution as seeking a degree or formal award (Data Cookbook – Common Data Set, n.d.). High school students also enrolled in postsecondary courses for credit are not considered degree/certificate-seeking (NCES, n.d).

Enroll: A student is considered enrolled at the institution if they are registered in at least one course for the term in question (Data Cookbook – Central Washington University [CWU], n.d.).

Enrollment Management: "An organizational concept and a systematic set of activities designed to enable education institutions to exert more influence over their student enrollments. Organized by strategic planning and supported by institutional research, enrollment management activities concern student college choice, transition to college, student attrition and retention, and student outcomes. These processes are studied to guide institutional practices in the areas of new student recruitment and financial aid, student support services, curriculum development and other academic areas that affect enrollments, student persistence and student outcomes. For college" (Hossler and Bean, 2012).

Enrollment yield: The number of admitted students who actually enroll in at least one course. Usually presented as a ratio or percentage of the whole (Steinberg, 2010).

ETL: Extract, Transform Load (ETL) "is the process of extraction, transformation and loading during database use, but particularly during data storage use. It includes the following sub-processes: Retrieving data from external data storage or transmission sources; transforming data into an understandable format, where data is typically stored together with an error detection and correction code to meet operational needs; and transmitting and loading data to the receiving end (Janssen, n.d.).

First Year: A matriculated student who is in his/her first year of attendance at the university, who have not attended (or attempted course credit at) another university after high school graduation. Being classified in this category is without regard to summer term credits and/or transfer credits earned through Running Start, Cornerstone or another dual-credit high-school/college program. This category of student includes admit types: FYR (First year), FYT (First year transfer), and IFY (International first year) (Data Cookbook – CWU, n.d.).

FYR: A first year student who enters with no transfer credits (Data Cookbook – CWU, n.d.).

FYT: A first year student who enters with transfer credits earned through Running Start, Cornerstone, dual credit high-school/college program, AP exam or International Baccalaureate (Data Cookbook – CWU, n.d.).

Graduate: Attainment of a pre-specified degree upon completion of all degree requirements.

Graduation rate: The rate required for disclosure and/or reporting purposes under Student-Right-to-Know Act. This rate is calculated as the total number of completers within 150% of the normal time divided by the revised adjusted cohort (NCES, n.d.).

Hispanic or Latino: A person of Cuban, Mexican, Puerto Rican, South or Central American, or other Spanish culture or origin, regardless of race (NCES, n.d).

Institution of higher education: a term formerly used in IPEDS and HEGIS to define an institution that was accredited at the college level by an agency or association recognized by the Secretary, U.S. Department of Education. These schools offered at least a one-year program of study creditable toward a degree and were eligible for participation in Title IV Federal financial aid programs Graduate: Attainment of a pre-specified degree upon completion of all degree requirements (NCES, n.d.).

Native Hawaiian or Other Pacific Islander: A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands (NCES, n.d.).

OLAP: online analytical processing enables (OLAP), a category of software tools that provides analysis of data stored in a database. OLAP tools enable users to analyze different dimensions of multidimensional data (Webopedia, n.d.).

Persist/Persistence: "The act or fact of persisting" (Merriam-Webster, n.d.). For the purpose of this paper, the author differentiates between persistence and retention as follows: Persistence or to persist indicates that the student continues enrollment through

completion of degree attainment; whereas, retention indicates that the student returned the following academic year – usually fall to fall.

Program: Also known as Academic Program. An instructional program leading toward an associates, bachelors, masters, doctors, or first-professional degree or resulting in credits that can be applied to one of those degrees (NCES, n.d)

Predictive Analytics: Technology that learns from experience (data) to predict the behavior of individuals in order to drive better decisions (Siegle, 2013, p. 107). Involves extracting data from existing data sets with the goal of identifying trends and patters. These trends and patterns are then used to predict future outcomes and trends. While it's not an absolute science, predictive analytics does provide companies with the ability to reliably forecast future trends and behaviors (NG Data, n.d., para 1)

Predictive Model: "A mechanism that predicts a behavior of an individual, such as click, buy, lie , or die. It takes characteristics of the individual as input, and provides a predictive score as output. The higher the score, the more likely it is that the individual will exhibit the predicted behavior" (Siegel, 2013, p. 25).

Public Institution: An educational institution whose programs and activities are operated by publicly elected or appointed school officials and which is supported primarily by public funds (NCES, n.d.).

Race / Ethnicity: Categories developed in 1997 by the Office of Management and Budget (OMB) that are used to describe groups to which individuals belong, identify with, or belong in the eyes of the community. The categories do not denote scientific

definitions of anthropological origins. They are used to categorize US Citizens, resident aliens and other eligible non-citizens. Individuals are asked to first designate ethnicity as: Hispanic or Latino or Not Hispanic or Latino. Second, individuals are asked to indicate one or more races that apply among the following: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White (NCES, n.d).

Race and ethnicity unknown: The category used to report students or employees whose race and ethnicity are not known(NCES, n.d.).

Recruit: to attempt to enroll or enlist (a member, affiliate, student or the like) (recruit, n.d.).

Retention: Undergraduate degree seeking students who enroll consecutively from one academic year to the next academic year (Data Cookbook – CWU, n.d.).

Retention rate: Used for reporting purposes - A measure of the rate at which students persist in their educational program at an institution, expressed as a percentage. For four-year institutions, this is the percentage of first-time bachelors (or equivalent) degree-seeking undergraduates from the previous fall who are again enrolled in the current fall. For all other institutions this is the percentage of first time degree/certificate-seeking students from the previous fall who either re-enrolled or successfully completed their program by the current fall (NCES, n.d.).

Stop out: A student who left the institution and returned at a later date(NCES, n.d).

Student success: Provides educationally purposeful programs, events services and activities that promote academic, personal and professional growth within and beyond the classroom (Central Washington University, n.d.).

Student support services: Any program, service, offering, action, intervention or policy at an institution that supports or assists students in the successful completion of a given course and/or completion of degree or credential of value in the workplace. These supports can be provided to students pursuing an on-ground or online education and can be delivered via a range of modalities. Supports can be proactive, aimed at preventing issues before they start, for example good advising for a program major that stimulates the student, or reactive and necessarily aimed at addressing issues that arise, including alerts for students who don't turn in a first assignment or special programs for those on academic probation. Supports can be directly related to a specific student's academic course work (example: tutoring) or they can be part of the overall academic infrastructure to promote student success (example: required office hours by instructors; clubs and organizations that enhance social integration) (Data Cookbook – Predictive Analytics Reporting Framework, n.d.).

Transfer credit: The policies and procedures used to determine the extent to which educational experiences or courses undertaken by a student while attending another institution may be counted for credit at the current institution (NCES, n.d).

Transfer student: A student entering the reporting institution for the first time but known to have previously attended a post secondary institution at the same level (e.g.

undergraduate, graduate). This includes new students enrolled n the fall term who transferred into the institution the prior summer term. The student may transfer with or without credit (NCES, n.d.).

Tuition: The amount of money charged to students for instructional services. Tuition may be charged per term, per course, or per credit (NCES, n.d.).

Undergraduate: A student enrolled in a 4- or 5- year bachelor's degree program, an associate's degree program, or a vocational or technical program below the baccalaureate (NCES, n.d.).

White: A person having origins in any of the original peoples of Europe, the Middle East, or North Africa (NCES, n.d.).

Appendix B: Decision Tree Rules, Coding, Comparisons

First Year Decision Tree Rules:

DT Rules/Nodes
Yes- 3, 91, 361, 47, 181,
No- 360, 44, 10, 46, 4

Excel code to identify DT nodes.
=IF(D2>=3.311,3,IF(AND(D2<3.311,H2=0,I2<1556,M2<0.5,D2>=2.792,E2>=3.21),91,
IF(AND(D2<3.311,H2=0,I2<1556,M2<0.5,D2>=2.792,E2<3.21,K2<0.5,L2>=0.5),361,I
F(AND(D2<3.311,H2=0,I2<1556,M2>=0.5,C2=0),47,IF(AND(D2<3.311,H2=0,I2<1556
,M2<0.5,D2>=2.792,E2<3.21,K2>=0.5),181,IF(AND(D2<3.311,H2=0,I2<1556,M2<0.5,
D2>=2.792,E2<3.21,K2<0.5,L2<0.5),360,IF(AND(D2<3.311,H2=0,I2<1556,M2<0.5,D2
<2.792),44,IF(AND(D2<3.311,H2=0,I2>=1556),10,IF(AND(D2<3.311,H2=0,I2<1556,
M2>=0.5,C2=1),46,IF(AND(D2<3.311,H2=1),4,"OTHER")))))))))

Tree as rules:

Rule number: 3 [DEGREE_COMPLETION_FLAG=Yes cover=1793 (36%) prob=0.69]
HIGH_SCHOOL_GPA>=3.311
Actual: 2535 OF 7077 OR 35.82%

Rule number: 91 [DEGREE_COMPLETION_FLAG=Yes cover=64 (1%) prob=0.67]
HIGH_SCHOOL_GPA< 3.311
DEVELOPMENTAL_MATH_FLAG=No
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 1556
TIN_GENDER_F< 0.5
HIGH_SCHOOL_GPA>=2.792
MAX_GPA_FROM_EXTERNAL_INSTITUTION>=3.21
Actual: 93 OF 7077 OR 1.31%

Rule number: 361 [DEGREE_COMPLETION_FLAG=Yes cover=42 (1%) prob=0.64]
HIGH_SCHOOL_GPA< 3.311
DEVELOPMENTAL_MATH_FLAG=No
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 1556
TIN_GENDER_F< 0.5
HIGH_SCHOOL_GPA>=2.792
MAX_GPA_FROM_EXTERNAL_INSTITUTION< 3.21
TIN_Race_Ethnicity_European.Middle.Eastern.White< 0.5

TIN_Race_Ethnicity_Latino.Hispanic>=0.5
Actual: 59 OF 7077 OR .83%

Rule number: 47 [DEGREE_COMPLETION_FLAG=Yes cover=842 (17%) prob=0.60]
HIGH_SCHOOL_GPA< 3.311
DEVELOPMENTAL_MATH_FLAG=No
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 1556
TIN_GENDER_F>=0.5
FIRST_GENERATION_FLAG=No
Actual: 1207 OF 7077 OR 17.06%

Rule number: 181 [DEGREE_COMPLETION_FLAG=Yes cover=451 (9%) prob=0.54]
HIGH_SCHOOL_GPA< 3.311
DEVELOPMENTAL_MATH_FLAG=No
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 1556
TIN_GENDER_F< 0.5
HIGH_SCHOOL_GPA>=2.792
MAX_GPA_FROM_EXTERNAL_INSTITUTION< 3.21
TIN_Race_Ethnicity_European.Middle.Eastern.White>=0.5
Actual: 662 OF 7077 OR 9.35%

Rule number: 360 [DEGREE_COMPLETION_FLAG=No cover=259 (5%) prob=0.42]
HIGH_SCHOOL_GPA< 3.311
DEVELOPMENTAL_MATH_FLAG=No
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 1556
TIN_GENDER_F< 0.5
HIGH_SCHOOL_GPA>=2.792
MAX_GPA_FROM_EXTERNAL_INSTITUTION< 3.21
TIN_Race_Ethnicity_European.Middle.Eastern.White< 0.5
TIN_Race_Ethnicity_Latino.Hispanic< 0.5
Actual: 357 OF 7077 OR 5.04%

Rule number: 44 [DEGREE_COMPLETION_FLAG=No cover=386 (8%) prob=0.39]
HIGH_SCHOOL_GPA< 3.311
DEVELOPMENTAL_MATH_FLAG=No
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 1556
TIN_GENDER_F< 0.5
HIGH_SCHOOL_GPA< 2.792
Actual: 551 OF 7077 OR 7.79%

Rule number: 10 [DEGREE_COMPLETION_FLAG=No cover=958 (19%) prob=0.36]

HIGH_SCHOOL_GPA< 3.311
DEVELOPMENTAL_MATH_FLAG=No
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT>=1556
Actual: 1379 OF 7077 OR 19.49

Rule number: 46 [DEGREE_COMPLETION_FLAG=No cover=76 (2%) prob=0.36]
HIGH_SCHOOL_GPA< 3.311
DEVELOPMENTAL_MATH_FLAG=No
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 1556
TIN_GENDER_F>=0.5
FIRST_GENERATION_FLAG=Yes
Actual: 110 OF 7077 OR 1.55%

Rule number: 4 [DEGREE_COMPLETION_FLAG=No cover=84 (2%) prob=0.00]
HIGH_SCHOOL_GPA< 3.311
DEVELOPMENTAL_MATH_FLAG=Yes
Actual: 124 OF 7077 OR 1.75%

[1] 19 15 13 18 16 1 14 9 6 10 7 4 2 11 12 8 5 17 3

Generated by RStat 2015-07-14 15:21:58

First Year Students of Color Decision Tree Rules, Coding, Comparisons:

Decision Tree Rules/Nodes:
Yes – 231, 15, 27, 59, 461
No - 460, 26, 58, 114, 12, 56, 2

=IF(AND(J2=0,D2>=2.779,D2<3.485,H2>=357,AD2>=815,M2>=0.5,F2>=9),231,
IF(AND(J2=0,D2>=2.779,D2>=3.485),15,IF(AND(J2=0,D2<2.779,H2<566.5,K2>=0.5),
27,IF(AND(J2=0,D2>=2.779, D2<3.485, H2<357,I2=0),59,IF(AND(J2=0, D2>=2.779,
D2<3.485, H2>=357, AD2>=815, M2>=0.5,F2<9,L2<0.5),461,
IF(AND(J2=0,D2>=2.779, D2<3.485,
H2>=357,AD2>=815,M2>=0.5,F2<9,L2>=0.5),460,IF(AND(J2=0,D2<2.779,
H2<566.5,K2<0.5),26,IF(AND(J2=0,D2>=2.779, D2<3.485,H2<357,
I2=1),58,IF(AND(J2=0,D2>=2.779,D2<3.485,H2>=357,AD2>=815,M2<0.5),114,IF(AN
D(J2=0,D2<2.779, H2>=556.5),12,IF(AND(J2=0,D2>=2.779, D2<3.485, H2>=357,
AD2<815),56,IF(J2=1,2,"other")))))))))))

Tree as rules:

Rule number: 231 [DEGREE_COMPLETION_FLAG=Yes cover=18 (2%) prob=0.89]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA>=2.779
HS.GPA< 3.485
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT>=357
HIGHEST_SATACT_COMPOSITE>=815
TIN_ORIGINAL_ADMIT_TYPE_CODE_FYR>=0.5
TRANSFER_CREDITS_ACCEPTED>=9
25 of 1142 or 2.19%

Rule number: 15 [DEGREE_COMPLETION_FLAG=Yes cover=142 (18%) prob=0.68]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA>=2.779
HS.GPA>=3.485
204 of 1142 or 17.86

Rule number: 27 [DEGREE_COMPLETION_FLAG=Yes cover=34 (4%) prob=0.65]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA< 2.779
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 566.5
TIN_ETHNICITY_Latino.Hispanic>=0.5
41 of 1142 or 3.6%

Rule number: 59 [DEGREE_COMPLETION_FLAG=Yes cover=242 (30%) prob=0.64]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA>=2.779
HS.GPA< 3.485
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 357
FIRST_GENERATION_FLAG=No
348 of 1142 or 30.47%

Rule number: 461 [DEGREE_COMPLETION_FLAG=Yes cover=41 (5%) prob=0.63]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA>=2.779
HS.GPA< 3.485
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT>=357
HIGHEST_SATACT_COMPOSITE>=815
TIN_ORIGINAL_ADMIT_TYPE_CODE_FYR>=0.5
TRANSFER_CREDITS_ACCEPTED< 9
TIN_ETHNICITY_Not.Latino.Hispanic< 0.5

67 of 1142 or 5.87%

Rule number: 460 [DEGREE_COMPLETION_FLAG=No cover=65 (8%) prob=0.42]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA>=2.779
HS.GPA< 3.485
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT>=357
HIGHEST_SATACT_COMPOSITE>=815
TIN_ORIGINAL_ADMIT_TYPE_CODE_FYR>=0.5
TRANSFER_CREDITS_ACCEPTED< 9
TIN_ETHNICITY_Not.Latino.Hispanic>=0.5
91 of 1142 or 7.97%

Rule number: 26 [DEGREE_COMPLETION_FLAG=No cover=60 (8%) prob=0.40]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA< 2.779
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 566.5
TIN_ETHNICITY_Latino.Hispanic< 0.5
86 of 1142 or 7.53%

Rule number: 58 [DEGREE_COMPLETION_FLAG=No cover=30 (4%) prob=0.33]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA>=2.779
HS.GPA< 3.485
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT< 357
FIRST_GENERATION_FLAG=Yes
39 of 1142 or 3.42%

Rule number: 114 [DEGREE_COMPLETION_FLAG=No cover=30 (4%) prob=0.27]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA>=2.779
HS.GPA< 3.485
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT>=357
HIGHEST_SATACT_COMPOSITE>=815
TIN_ORIGINAL_ADMIT_TYPE_CODE_FYR< 0.5
43 of 1142 or 3.77%

Rule number: 12 [DEGREE_COMPLETION_FLAG=No cover=79 (10%) prob=0.24]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA< 2.779
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT>=566.5

117 of 1142 or 10.25%

Rule number: 56 [DEGREE_COMPLETION_FLAG=No cover=36 (4%) prob=0.17]
DEVELOPMENTAL_MATH_FLAG=No
HS.GPA>=2.779
HS.GPA< 3.485
FIRST_QUARTER_FINANCIAL_AID_DISBURSED_AMOUNT>=357
HIGHEST_SATACT_COMPOSITE< 815
54 of 1142 or 4.73%

Rule number: 2 [DEGREE_COMPLETION_FLAG=No cover=23 (3%) prob=0.00]
DEVELOPMENTAL_MATH_FLAG=Yes
27 of 1142 or 2.36%

[1] 19 23 8 22 18 20 9 15 10 3 1 16 13 6 11 17 7 4 21 14 5 12 2

Generated by RStat 2015-07-17 08:56:17