DATA MINING TECHNIQUES AND MATHEMATICAL MODELS FOR THE OPTIMAL SCHOLARSHIP ALLOCATION PROBLEM FOR A STATE UNIVERSITY

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

by

SHUAI WANG
B.S., Management, Dalian Jiaotong University, 2011

2017
Wright State University

Wright State University
GRADUATE SCHOOL

December 14, 2017

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Shuai Wang ENTITLED Data Mining Techniques and Mathematical Models for the Optimal Scholarship Allocation Problem for a State University BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Xinhui Zhang, Ph.D.
Dissertation Director

Frank W. Ciarallo, Ph.D.
Director, Ph.D. in Engineering Program

Barry Milligan, Ph.D.
Interim Dean of the Graduate School

Committee on
Final Examination

Xinhui Zhang, Ph.D.

Pratik Parikh, Ph.D.

Caroline Cao, Ph.D.

Subhashini Ganapathy, Ph.D.

Nan Kong, Ph.D.

ABSTRACT

Wang, Shuai. Ph.D in Engineering Program, Department of Biomedical, Industrial and Human Factors Engineering, Wright State University, 2017. *Data Mining Techniques and Mathematical Models for the Optimal Scholarship Allocation Problem for a State University*.

Enrollment Management and Financial Aid. Enrollment management is the term that is often used to describe the synergistic approaches to influence the enrollment of higher education institutions, and consists of activities such as student college choice, transition to college, retention, and graduation. Of all the factors, financial aid, institution rank, and tuition are the three most important ones that affect students' choice processes and matriculation decisions; as such, with the continuous increase of tuition over the years, financial aid serves as a marketing tool and plays an important role in attracting students. In the United States, in the 2012-2013 academic year, there were a total of 20.4 million students enrolled in degree-granting institutions and more than eighty percent of them received financial.

The Optimal Scholarship Allocation Problem: The widespread use of financial aid leads to an important problem yet to be solved in the literature, i.e., how to optimally allocate the limited financial aid to students with various social and economic backgrounds so as to achieve enrollment goals. Though financial aid can be of various forms, merit-based scholarships are the primary part of the allocation process. This problem, referred to as *the optimal scholarship allocation problem*, has puzzled the enrollment management teams at many higher institutions and is the focus of this thesis.

Solution Approach: This thesis proposes a series of predictive and optimization models to solve the optimal financial aid allocation problems. The methodology consists of three sequential phases: 1) predictive models to find the responses (enrollment and graduation probabilities and years of study) to various levels of scholarship for students with various socioeconomic backgrounds; 2) optimization models to find the maximum revenue for given budget based on the response discovered to the various levels of scholarships; and 3) data mining models to discover patterns and transform results from the optimization model to simple and effective policies.

Phase I: Predictive Models. A series of predictive models have been investigated to estimate the responses from students to various levels of scholarship awards. These responses can be classified into two categories: the first category includes enrollment and graduation decisions and the second one is the number of years of study once a student enrolls in the institution. In the first category, because of the binary nature of the responses (enroll or not enroll), logistic regression based models have been adopted to predict the probability of enrollment and the probability of graduation given that student enrolls. In the second category, regression analysis are adopted.

Phase II: Optimization Models. An optimization model is designed to allocate financial aid to applicants with an objective to maximize the revenue, which is composed of net tuition, i.e., tuition minus scholarship, over the years of study, plus the state share of instruction once the student graduates. The constraints to be observed include the total budget limitations and a fairness constraint. For a merit-based scholarship, the fairness constraint stipulates that a student with better academic performance must be assigned to an equal or higher level of scholarships than that of students with a lower academic

performance. The inclusion of the fairness constraint has dramatically increased the size of the model, and to reduce computational burden, the concept of a minimum dominance set is developed. This has reduced the size of the model by orders of magnitude and enabled the efficient solution of the resulting mathematical model.

Phase III: Policies Analysis Models. Regression analysis is developed to discover patterns in the optimization results, in the form of the amount of scholarship awarded for each student, and translate them into simple and effective scholarship award policies for implementation. Several techniques such as decision tree and piecewise regression have been explored. For the institution under study, the results suggested that a composite score based on the student's GPA and ACT scores can be used as the basis for the award of scholarships; and a simple yet effective award scholarship policy derived from piecewise regression has been discovered.

Implementation: The analysis based on the above framework was adopted by the institution under study and has been used in an overhaul of the scholarship redesign. The piecewise regression derived, composite score based scholarship award policy proves to be effective, and together with a proactive marketing strategy it has yielded an 11% increase in directly admitted students under a similar budget. This translates into millions of dollars of revenue and significantly improves the university's bottom line.

Contents

List of Figures

List of Tables

Acknowledgment

I would like to take this opportunity to extend my thanks to my dissertation committee, Dr. Xinhui Zhang, Dr. Pratik Parikh, Dr. Caroline Cao, Dr. Subhashini Ganapathy, and Dr. Nan Kong, for taking their time and effort in serving in the committee. Thank you for the invaluable suggestions to my study. I would like to express my deepest gratitude to my advisor Dr. Zhang for his encouragement, patience and knowledge.

I would like to thank the Kroger Operations Research team for providing various opportunities and projects to sponsor my Ph.D. study. These projects involve various aspects of operations research, from the integrated regression and time series based forecasting and inventory management, periodic vehicle routing, collaborative category optimization, and staff scheduling optimization. These projects opened my eyes and helped me to mature as an operations researcher.

I would like to thank my friends from the BIE department and the large scale optimization lab (Lab 249): Dr. Yan Liu, Dr. Qiang Liu, Dr. Gamze Kilincli, Dr. Lebin Lin, Isaac Hampton, Hakan Gecili, Lijian Xiao and Jue Huang. Thanks for the fun time that we spent together. I would also thanks my friends from the computer science department: Zhongliang Li, Dr. Ming Tan, and Dr. Shaodan Zhai.

Dedicated to

My Parents and My Family

Introduction

1.1 Enrollment Management and Financial Aid

Enrollment and Financial Aid. Enrollment management was coined by Dr. Jack Maguire (Maguire, 1976) and is the term that is often used to describe the synergistic approach to shape the enrollment of institutions to meet established goals, such as to increase the number of high-caliber students, to diversify the student body, and to retain more students (Kemerer et al., 1982). Enrollment management consists serious of activities such as student college choice, transition to college, retention, and graduation (Hossler and Bean, 1990) and is critical to many colleges and universities (Braunstein et al., 1999; Maltz et al., 2007; Aksenova et al., 2006).

Financial aid is an integral part of enrollment management strategies for institutions (Dynarski and Scott-Clayton, 2013). Financial aid is funding provided to students to cover various costs such as tuition, fees and board while they attend an institution. Financial aid comes from federal and state programs, as well as private institutions and agencies. Financial aid can be awarded in different forms such as grants, education loans, and scholarships. Grants are money that students do not have to repay, such as the federal Pell Grants. Loans

such as Free Application for Federal Student Aid (FAFSA) are money students borrow and must pay back, and which usually carry interest. Scholarships are given to students based on desired qualities such as academic achievement, athletic ability.

Purpose of Financial Aid. Financial aid serves multiple purposes, such as providing access and affordability to families who need financial help, stimulating more students to major in areas having labor shortages; moreover, financial aid programs have served the purpose of a marketing tool for institutions to attract students and shape its enrollment.

From early studies, Heller (1997) and Leslie and Brinkman (1988) noticed that "receiving a financial aid award has a significant positive effect on the likelihood that a student will enter the institution that has made the financial aid offer" and "the effect of just receiving an award, regardless of the amount, equals or exceeds the effects of the amount of the award." Leslie and Brinkman (1987), in an early review on the relationship between price and enrollment in higher education, suggested that "higher prices reduce higher education enrollments, or as tuition increases, many high school students can not afford college and enrollment might decrease; however, this has not been observed in practice and a large number of students do attend college." and "the above quandary and the answer was partially due to the ameliorating effects of financial aid".

In fact, various studies suggest that of all the factors, financial aid, institution rank, and tuition are the three most important ones that affect students' choice processes and matriculation decisions (Fuller, 2014); as such, with the continuous increase of tuition over the years, financial aid plays an even more important part to price discriminate potential applicants and is widespread among many institutions.

Ubiquity of Financial Aid. According to U.S. Department of Education (National Center

for Education Statistics, 2014), in the 2012-2013 academic year, all degree-granting institutions had total revenues of over $554 billion, with over $280 billion from student tuition and fees. The average undergraduate tuitions and fees were $15,640 and $35,987 for public and private school respectively. There were a total of about 20.4 million degree-seeking students; 84.4% of students received some types of aid, 72.4% received grants and 56.7% received loans.

1.2 The Scholarship Allocation Problem

The Optimal Scholarship Allocation Problem. The widespread use of financial aid leads to an important problem yet to be solved in the literature, i.e., how to optimally allocate the limited financial aid to students with various social and economic backgrounds so as to achieve enrollment goals. Though financial aid can be of various forms, merit-based scholarships are the primary part of the allocation process. This problem, referred to as *the optimal scholarship allocation problem*, has puzzled the enrollment management teams at many higher institutions and is the focus of this thesis.

Gap in the Related Literature. It bears mentioning that in the past few decades many studies have been conducted to evaluate the impact of changes in tuition and financial aid on students' enrollment decisions in higher education. These studies can be classified into two categories, macro-level student demand studies and micro-level student choice models.

Macro-Level Student Demand Studies. Most research studies of financial aid are macro-

level student demand studies or the use of market level data for analyzing on the effects of tuition and financial aid on students' decisions. For example, Hossler et al. (1989) have consistently found that "African American students and Latino students are more cost sensitive and more responsive to financial aid offers than students of similar socioeconomic background". Braunstein et al. (1999) found that for every $1,000 increase in financial aid, the probability of enrollment increased between 1.1% and 2.5%. These studies have been studied the effectiveness of various federal and state financial aid programs, such as the HOPE program, the CalGrant program, and the Adams program. Most of these studies, however, only address the effects of financial aid on enrollment over a longitude of years across institutions and are not intended to address the optimal allocation of scholarship for each institution.

Micro-Level Student Choice Process and Target Market Optimization. It is accepted that the response to the changes in financial aid could differ among various groups of students. For example, changes in tuition and financial aid affect poorer students more than wealthier; changes in financial aid affect minority students more than white students. As students have diverse social and economic backgrounds, these studies suggest the response to financial aid differs from student to student, and the study of these factors and the decision process is referred to as the college choice process model (Paulsen, 1990). For example, Jackson (1978) created a general model of students' postsecondary decision processes as a function of place, background, school, student, friends, occupation, aspiration, plans, colleges and jobs.

Given the various factors in the college choice process, it would be desirable to predict the students' responses to the enrollment decision at specific university. In many respects,

this problem is similar to problems addressed in the targeted marketing literature such as (Reinartz and Kumar, 2003). Although the enrollment study shares some similarities with the marketing targeting study, the study of the response of individual students and the optimal selection of the target set of students can still be rather complex to solve.

Complexity and Importance of the Optimal Scholarship Allocation Problem. The optimal scholarship allocation problem, however, is more complicated due to several reasons.

First and foremost, there exist many factors that affect students' college choice and enrollment decisions. For example, students' college decision are affected by their own aspirations, family background, et cetera, and not all these factors are available to researchers and let alone to the decision makers.

Second, it is difficult to construct models that separate the impact of tuition from the effects of financial aid, and as of now, different practices exist in higher education to balance tuition and financial aid. For example, some public institutions apply low tuition and offer little financial aid, and rely on low tuition as the primary financial enticement; yet other institutions could pursue a high tuition, high financial aid (the *Robin Hood* strategy), to achieve enrollment goals.

Third, there are not as many optimization studies on the allocation of financial aid as have been seen in other industries such as airlines, thus there has not been any guidance in the solution of these problems.

It is hypothesized, and will be demonstrated in this thesis, however, that it is possible for individual institutions to study the impact of financial aid on students in the application pool, and moreover to optimize the allocation of financial aid to increase its enrollment

goals.

The solution to the optimal scholarship decision problem could have a direct and significant impact on the institution's financial health: on the one hand, tuition is an important part of an institution's revenue, and excessive use of financial aid could potentially reduce its revenue; on the other hand, insufficient use of financial aid could potentially reduce student enrollment and thus undermine the total revenue.

The Dilemma at a State University. This study was motivated by requirements from the executive teams at a public state university to study the effectiveness of its financial aid policies and potentially optimize its allocation to boost enrollment and increase its financial bottom line. The state university is one of the thirteen state universities within Ohio and had an enrollment of 17,779 in the 2013-2014 academic year. At the time of the study, the university aims to provide an affordable yet high-quality education experience and is eager to grow its enrollment and revenue in the coming years. The desire to grow, however, is faced with tough challenges because the university is competing with other flagship state universities for high-talented students and with community colleges for students seeking affordability.

Raising tuition, and accordingly allocating a portion of the increase in financial aid to target students, is not an option for the state university. To provide affordable education, the state government has capped statewide tuition increases, and has mandated the increase of tuition to be zero for the 2015 - 2016 year. More than 95% of the students, whose tuition contributes to 46.75% of the university's revenue, come from the surrounding region where the population remains steady with no dramatic increase in high school graduates. As such, the effective use of institutional resources such as financial aid to increase enrollment has

been the focus of university executives. Though several questions have been raised, these questions boil down to the solution of the financial aid allocation problem, i.e., what levels of scholarship students should be awarded to maximize the net revenue for the institution.

1.3 A Three Phase Solution Approach

In the solution of the optimal financial aid decision problem, it is necessary to solve several sequential problems related to a) the determination of responses to financial aid award from students with various demographic, academic, and financial backgrounds, b) the allocation of the scholarship to find the maximum revenue for a given budget based on the response discovered to the various levels of scholarships, and c) derivation of discovered patterns and transform of the results from the optimization model into simple and effective scholarship award policies .

Various techniques have been investigated in the three-phase framework in the solution of the optimal scholarship allocation problem and policy analysis and the details of the techniques and models developed in this research are summarized presently.

The solution to these problems form a three-phase methodologies proposed in this study as follows:

Phase I: Predictive Models. One of the main objectives of the optimal scholarship assignment problem to maximize the revenue, which in this study is composed of two parts. The first one is the net tuition income from students, which is the difference between tuition and fees, over the years of the studies at the institution. The second one is the state share of instruction, known as SSI, which is awarded to the institution once the students gradu-

ate. As such, it is necessary to derive analytic models to predict the probability of student matriculation decisions, to predict the number of years of study at the institution, and to predict the probability of graduation from the institution. Due to dropouts and transfers, not all students will graduate from the institution.

To this end, a series of predictive models have been investigated to estimate the two categories of responses from students to various levels of scholarship awards. The first category is the enrollment and graduation decisions and the second one is the number of years of study once a student enrolls in the institution. In the first category, because of the binary nature of the responses ("enroll" or "not enroll"), logistic regression based models have been adopted to predict the probability of enrollment and the probability of graduation given that student enrolls. In the second case, regression analysis are adopted.
Phase II: Optimization Models. An integer linear model is designed to optimally allocate financial aid to applicants with an objective to maximize the revenue. The optimal problem by itself is subject to two constraints: the first one is the budget limitation on the total scholarship and the second one is the fairness of the award, which states that a student with higher academic performance must be be awarded an equal or higher level of scholarship than a student with lower academic performance. These two constraints make it impossible to solve the scholarship assignment models at individual levels, and the large number of pairwise fairness constraints requires the development of a customized algorithm to effectively solve the problem at the aggregate budget level.

To reduce computational burden, the concept of a minimum cardinality dominance set is developed, which has reduced the size of the model by orders of magnitude and enabled the efficient solution of the resulting mathematical model. Computation experiments show

that the use of minimum cardinality dominance has achieved a dramatic reduction regarding model size. In a test case with pairwise comparison of more than 5,200 students, this reduced over 13.8 million constraints to only 191,000 constraints, which enables effective solution of the models. In this particular case, the original model is computationally unsolvable, actually running out of memory; despite the large model size, the reduced model can be solved in only minutes.

Phase III: Policies Analysis Models. Regression analysis is developed to discover patterns in the optimization results, in the form of the amount of scholarship awarded for each student, and translate them into simple and effective scholarship award policies for implementation. Several techniques such as decision tree and piecewise regression have been explored. For the institution under study, the results pointed to the use of a composite score based on the student's GPA and ACT scores as a practicable basis for the award of scholarships and thus a simple yet effective award scholarship policy derived from piecewise regression has been discovered.

1.4 Implementation and Financial Results

The methodology based on the above framework was adopted by the institution under study and its use resulted in an overhaul of the scholarship design. The piecewise regression derived and composite score based scholarship award policy was used as the foundation for the scholarship award for the university in the 2013 to 2014 academic year. A proactive marketing approach has been taken where the enrollment and admission office has obtained data on student performance from ACT and potential students are awarded

the scholarship before they even apply. This has yielded an significant increase in directly admitted students under a similar budget.

Table 1.1 presents the enrollment statistics for the university in the 2012 - 2013 and 2013 - 2014 academic year after the new policy was applied:

	2013	2014	# Increase	% Increase
Application	6,101	6,068	-43	-0.7%
Admitted	4,541	4,773	232	5.1%
Non-Scholarship	2,166	2,157	-9	-0.4%
Scholarship Award	2,375	2,616	241	10.1%
Matriculated	2,001	2,222	221	11.0%

Table 1.1: Comparison of enrollment between 2012-2013 and 2013-2014 Years

In the 2012-2013 academic year, there were a total of 6,101 applicants, of them, 4,541 were admitted. 2,375 were awarded scholarships and 2,166 were not awarded scholarships. 52% of the students were awarded scholarships and a total of 2,001 students matriculated. In the 2013-2014 academic year, there were a total of 6,068 applicants; of them, 4,773 were admitted, 2,616 were awarded scholarships and 2,157 were not awarded scholarships, 56% of the students are awarded scholarships and a total of 2,222 students matriculated.

Notice that the number of applicants does not change dramatically, actually showing a reduction of -43 (-0.7% increase), but the actual enrollment increased by 221 or 11.0% over that of the previous year. It is estimated that the use of the optimal policy could

generate millions of dollars of revenue for the university in the next few years.

1.5 Contribution and limitation

The research studies the optimal allocation of scholarships that faces the enrollment management of higher institutions. The problem is of significant importance to higher institutions as either over-spending or under-spending could negatively impact the institution's total net tuition revenue. The problem nevertheless has not been widely studied by the academic literature.

Contribution. This research proposes a set of analytic models to predict the students' response in terms of enrollment and graduation decisions to scholarship award, and an optimization model to determine the scholarship level. The methodologies elegantly integrate data mining (in the prediction of enrollment and graduation decisions and the derivation of award policies) and optimization techniques. The successful solution of these problems will fill the gap in the literature of the optimal financial aid allocation problem, contribute significantly to research studies in financial aid allocation, and has brought noticeable financial benefits to the university under the study.

Limitation. The prediction of student's enrollment decision is by itself a hard problem, because enrollment in a particular institution is determined by many factors that are not available to the institution. For example, it is well known that a student's enrollment decision is affected by family influence and aspiration level, which are not available to any institution without detailed survey studies. The research is also limited by the availability of data provided by the institution under study. Nevertheless, the methodology proposed in

this research could potentially be applied to similar college universities where merit-based aid could be an effective tool to attract students.

Organization of the Thesis. The remainder of the thesis is organized as follows. Section 2 presents a review of related literature. Section 3 presents the predictive models based on logistic regression on the probability of enrollment and graduation and Section 4 presents the predictive models for the number of years of study. Section 5 presents the mathematical models and techniques for computational improvement. Section 6 presents the policy development and implementation results, and finally Section 7 presents the conclusion and future research.

Literature Review

The solution of the financial aid allocation problems requires innovative models and techniques that draw insights from data mining and optimization techniques; as such, it is necessary to first look at traditional studies such as the macro-level student demand studies and the micro-level students college choice models and to understand thoroughly the underlying factors, financial aid being one of them, that affect the decision-making process of potential students.

2.1 Macro-Level Student Demand Models

Early enrollment research studies stemmed from economics and were prompted to answer the questions related to the effect of raising the price of education on enrollment decisions. These questions include "What happens to enrollment when colleges and universities raise their prices?" "Who, if anyone, is sent away?" "What is the net impact of higher prices and reduced enrollments upon institutional finances?" These questions lead to what are called "the student demand studies", as an outgrowth of demand theory in economics (Leslie and Brinkman, 1987, 1988; Heller, 1997; Ehrenber, 2004; Crouse, 2015).

Demand theory holds that the quantity of a particular good is a function of price, the income of the buyers, the prices of other goods, and the buyers' tastes or preferences. In the student demand studies, these factors relate to tuition, financial aid (viewed as a discount), income, race, student preferences et cetera.

2.1.1 Student Demand Theory on Tuition

Leslie and Brinkman (1987) presented an early review of literature on the relationship between price and enrollment in higher education. The authors reviewed 25 empirical student demand works in the 1980s. The results show that:

> higher prices reduce higher education enrollments, that students historically have been more responsive to tuition prices than to (offsetting) student aid, and that low-income students are most sensitive to price changes, as are students in public versus private institutions.

Many more studies have been released since then. These research studies reached similar conclusions: higher prices reduce higher education enrollments. For example, Leslie and Brinkman (1987) derived the calculation of a student price response coefficient (SPRC) and found that the mean price response is about -0.7%. "That is, for every $100 increase in tuition price, given year 1982 - 1983 average weighted higher education prices of $3,842 for tuition and room and board, one would expect an 18 - 24 year old participation rate drop of about three-quarters of a percentage point" (p,188).

The findings from these student demand studies with respect to tuition seem to suggest that higher prices reduce higher education enrollments, or as tuition increases, many

high school students can not afford college and enrollment might decrease; yet this has not been observed in practice and a large number of students do attend college over the period being examined. Leslie and Brinkman (1987) explained the quandary and the answer was partially due to the ameliorating effects of financial aid.

2.1.2 Student Demand Theory on Financial Aid

Leslie and Brinkman (1988), in a later study, reviewed 45 econometric analyses of relationships between student financial aid and college enrollment and pointed out that, i.e., receiving a financial aid has a significant positive effect on the student's decision to the institution. Heller (1997) provided a literature and summarized the results of these studies on the effects of different forms of financial aid , which are separate from the tuition.

Jackson (1988) used a cross-sectional analysis of the determinants of the demand for colleges. The results show that financial aid recipients were 6.5 percent and 7.8 percent more likely to enter colleges in 1972 and 1980 respectively than those who did not receive any financial aid. Braunstein et al. (1999) found that for every $1,000 increase in financial aid, the probability of enrollment increased between 1.1% and 2.5%. Crouse (2015) studied the nationwide tuition elasticity of for public two-year colleges and found that at the mean, a $100 increase in tuition resulted a decline in enrollment of about 0.883%.

Many universities are trying to find the balance between tuition income and financial aid spending. According to (Hossler et al., 1998), it was uncommon for schools to spend more than 10% to 15% of the tuition revenue as scholarship. However, private colleges now can have a high tuition and as high as 25% to 30% tuition discount. On the other hand, public schools use low tuition to attract students, and they are afraid that students

will be sensitive if tuition increases.

2.1.3 Target Effect on Financial Aid

Financial aid, nevertheless, is not simply a discount to the posted tuition price, but as Heller (1997) mentioned "a term that incorporates many different forms of student assistance, grants, loans and scholarships", for example, a 1000 dollar in loan is different than a 1000 dollar grant. So the interaction between financial aid and enrollment is not easy as it looks.

The effects on the increasing tuition and the *targeting effects* of financial aid to the students are different. For example, although all students are effected by the raising tuition, however, not all the students react the same to the financial aid: universities could offer extra targeted financial aid toward students with various background. For example, questions such as "do students from wealthier families have the same sensitivity to tuition increase compared to those from poorer families?", "Do white students react to financial aid awards similar to black students?" need to be answered.

- **Targeting Effect of Financial Aid on Students with Varying Characteristics**

Many of theses studies focused on how do students with different demographic and socioeconomic background react to the changes of tuition and financial aid (Jackson, 1978; Braunstein et al., 1999; Heller, 1997). A comprehensive review of the studies to answer these questions is not possible due to the sheer large volume; as such, only selected studies are listed below.

Income: "Lower income students are more sensitive to changes in tuition and aid than are

students from middle- and upper-income families" (Crouse, 2015).

Race: Research studies have consistently found that "African American students and Latino students are more cost sensitive and more responsive to financial aid offers than majority students of similar socioeconomic background" and "for Hispanic students, the results are mixed" Hossler et al. (1989).

Sectors: High-caliber students require more scholarship to be enrolled (Chapman and Jackson, 1987). The reason is that high-caliber students usually receive offers and scholarships from many universities and they tend to choose the universities with the highest reputation or highest scholarship. As a result, university has to offer a large scholarship to high-caliber students to complete with other university.

Unemployment Rate: Heller (1999) found that Asian, African American, and Hispanic students are more likely go to school when unemployment rate increases, which directly proves that bad economy spurs the interest of high education.

2.1.4 Student Demand Study for Policy Analysis

The student demand studies have been successfully applied to evaluate of the effectiveness of various programs such as Georgia's HOPE program, the CalGrant program, and the Massachusetts Adams Scholarship, to name a few.

Dynarski (2000) studied the impact of aid on the college attendance of middle- and upper- income youth by evaluating Georgia's HOPE scholarship and the results suggest a large impact on the attendance. They found that each $1,000 in aid increased the college

attendance rate in Georgia by 3.7 to 4.2 percent. The author also found evidence that the HOPE program actually widened the racial and income gaps in the enrollment.

Dynarski (2003) studied the impact of the elimination of the social security student benefit program. The author estimated that an offer of $1,000 in grant aid increases the probability of enrollment by about 3.6 percent.

Cohodes and Goodman (2014) applied a regression discontinuity design to study a Massachusetts merit aid program and the students' enrollment decisions and rates of degree completion. They found that students are willing to sacrifice college quality given relatively little financial aid, that students who made this decision were less likely to matriculate on average, diminishing the value of the extra enrollments.

Seftor and Turner (2002) found that the Pell Grant, the largest source of federal grants for college, has a positive effect on enrollment of potential students in their twenties and thirties. It has also been pointed out that though loans are the dominant form of aid today, little is known about how do they affect student behaviors (Dynarski and Scott-Clayton, 2013).

Abraham and Clark (2006) studied the District of Columbia's Tuition Assistance Grant Program (DCTAG) and concluded that under the program, the number of university applicants increases largely, especially the university which are eligible for the program subsidy. As a result, the actual enrollment of eligible universities increased. However, the overall enrollment in DC did not change very much and this implied that the DCTAG program has impact on where dose a student go rather than whether a student will go to college.

Castleman and Long (2016) found that the Florida Student Access Grant (FSAG) has a positive effect on the attendance at the eligible 4-year university as well as the degree completion rate.

Ehrenber (2004) surveys the discourse on the development of the econometrics of higher education over the last 40 years, and categorizes the surveyed articles accordingly. These categories include "the estimation of rates of return to higher education; determinants of college enrollment, college graduation, and choice of major; studies of the academic labor market; studies relating to models of university behavior; and studies relating to higher education as an industry and higher education governance".Finally, for some more recent reviews of the student demand studies, please see (Dynarski, 2002, 2003, 2000; Dynarski and Scott-Clayton, 2013).

Though financial aid can improve college access and completion, the actual enrollment and completion of college is not good as we assume due to the program delivery and other factors. Sjoquist and Winters (2015) found there is no statistical evidence on the effectiveness of certain state-based merit aid programs on college completion.

2.2 Enrollment Prediction at Micro-Level

The student demand studies investigate, at the macro level, enrollment decisions associated with students, yet it is still very hard to use these studies to evaluate the micro-level individual student's decision to enroll in particular schools. Carter and Curry (2011) pointed out that "published research using market-level data, though appropriate for national policy debates, is not necessarily useful for governance decisions at the university level".

2.2.1 College Choice Process and Models

The prediction of whether a student will attend a particular college is quite challenging. For example, it is well known that financial aid is one of the many variables that affect a students enrollment decision, but students may still turn down a full-ride if they are admitted by a prestigious college. As such, it is critical to understand the decision process and its related factors. These models are referred to as the college choice process models, which aim to address the following questions (Paulsen, 1990):

> (1) What factors are important to students of non-traditional age in making college decisions? (2) What are the phases of the college choice process? (3) What factors are important in creating a desire to attend college? (4) Why is the college search and application phase so important? (5) How can an institution more effectively manage enrollment in the selection and attendance phase?

Jackson (1978) derived a general model of students' post secondary decision processes as a function of ten variables, including background, aspiration, and friends. The author also stated that "the complexity of this model requires extensive attention if one wishes to weigh one background factor against another, or to determine which factors act upon the system and which act within".

Paulsen (1990) studied students' college choices to understand why students choose to attend one particular college over the other.

Micro-level enrollment model studies college choice behavior: estimate the probability of enrollment decision of individual student for a particular school.

2.2.2 Micro-level Response to Financial Aid and Its Optimization

Given the various factors in the college choice process, it would be desirable to predict a student's response to the enrollment decision at a university. In many respects, this problem is similar to the target marketing problem which are applicable to many other industries. One part of the targeting marketing problem is to make individual level offer decisions (Venkatesan and Kumar, 2004), and the problem of how does financial aid impact on the admission decision has many similarities with it. For example, similar to (Carter and Curry, 2011), this thesis uses an individual customer response model and proposes models to optimize the aid allocation.

Ehrenberg and Sherman (1984) provided one of the earliest optimization models to derive financial aid policies for a university. The authors used a single index (SAT scores) as the objective and proposed a model that allocates financial aid at the group level, with each member of a group receiving the same amount of offer.

Thanh and Haddawy (2007) proposed an approach to maximize tuition revenue through enrollment. The enrollment probability of each student is predicted by using a Bayesian network. An optimization model was developed to maximize tuition revenue subject to capacity and faculty-student ratio constraints. By adopting the optimization model, the institute can achieve the current enrollment level while reducing the financial budget.

Donald et al. (2010) studied how to combine empirical estimation of matriculation probability with optimal tuition pricing, which represents the optimal level of financial aid for each applicant, based on the demographic and academic information of applicants.

Sugrue (2010) used a constrained optimization technique to allocate merit-based aid at a medium-sized private university. The objective of the study was to attract higher quality students, measured by combined SAT score, so as to improve overall academic performance of enrolled students. The constraints of the study include the availability of students in certain SAT score ranges, the total budget available, and the enrollment limit.

Carter and Curry (2011) modeled individual student's choice and derived market level implications via upward aggregation to get college enrollment estimations. The authors are able to capture the real-time response of candidate schools of the student as well as the differentiating factors in the inter-university competition.

The paper proposed the use of tuition elasticities estimated by college and showed that "elastic demand can have deleterious effects on the quality of an incoming class even when demand for seats far outstrips supply".

Belloni et al. (2012) recently combines the optimal admission, scholarship decisions, and the choice of customized marketing offers to attract a desirable groups of students. The authors pointed out that this is a large targeted marketing and price discrimination problem which required a tailored approach to exploit the particular setting. The approach attempts to target a cohort of students based on an expected profile and then offer customized scholarship. The approach is tested in a field study of an MBA scholarship assignment process, and scholarship decisions were adjusted based on its results.

It bears mentioning that most of these studies tried to solve the college choice problem at the individual level, i.e. determine the value of the scholarship to a given student, but did not address the question that how to optimally allocate a scholarship budget at school level. What's more, these studies did not address the allocation of financial aid for students

with various socioeconomic characteristics, and thus did not fundamentally address the university's optimal scholarship allocation problems to reach institutional goals such as to maximize revenue. Though there have been a lot of mathematical models in various other industries, the use of mathematical models for enrollment management and financial aid allocation is rather limited.

2.3 Methodology Reviews

Statistical models have been widely used in these macro-level student demand studies and micro-level student choice and prediction models. For macro-level, the decision to be observed is the percentage of enrolled students, and cross-sectional statistical regression analysis is often used at the macro-level. For micro-level, however, the probability of "enroll" or "not enroll" is not directly observed, and only binary outcomes are observed. Logistic regression is typically adopted at the micro-level, as it is used to predict binary outcomes.

2.3.1 Regression Models in Student Demand Studies

Regression models have been largely applied to the analyze the impact of tuition and financial aid on students' decision. In mathematical way, Dynarski (2002) used the following equation to represent the relationship between the effect of financial aid and choice decisions: let S_i measuring student's decision such as matriculation, completion, and number of years stay in college; Aid_i be the amount of eligible aid for an individual; σ_i be unob-

served factors of decision, then the multivariate analysis can be stated as:

$$S_i = \alpha + \beta * Aid_i + \sigma_i$$

- **The Inclusion of Time Series Variables in the Model**

Hybrid approaches that combine multivariate analysis and time series can be seen in studies such as (Heller, 1999). In the examination of sources of variation in state spending on need-based aid, merit-based aid, and appropriations over the period 1990 to 2010, McLendon et al. (2014) used the lagged value of an outcome variable itself as a variable in a regression model to forecast the state spending on need-based and merit-based financial aid. Lavilles and Arcilla (2012) applied three types of time series models to forecast the number of students enrolled in a class.

- **The Inclusion of External Change Variables in the Model**

Dynarski (2003) argued that "the traditional approach to regress a person's educational attainment against covariates and the aid for which he is eligible and interpret the coefficient on aid as its causal effect" is problematic because of the complex nature of various characteristics which impact the matriculation decision. To identify the effect of financial aid, the authors suggest adding variation to the financial aid which is exogenous to unobservable factors that influence the matriculation. A similar approach was used in (Dynarski, 2003; Abraham and Clark, 2006).

Curs and Singell (2002) argued that application and enrollment decisions are correlated because the applicants from the pool share many similar features. Probit models were used to predict: first, the enrollment decision and second, the individual response to the net-price. The results show that in-state and out-of-state students react differently to the net-price changes (elasticity) and the differences are sensitive to both price variation exerted on both individuals and over time.

2.3.2 Logistic Regression Models in Student Choice Response Studies

Logistic regression, is a member of the generalized linear model used to solve binary classification problems. Independent variables in the logistic regression can be either categorical or continuous. Logistic regression has been successfully used in various marketing research studies; for details of these studies, please see (Hosmer et al., 2013).

Logistic regression has been used mostly for predicting enrollment levels and models the probability of enrollment as a linear function of a set of predictor variables (Peng et al., 2002). The predictor variables in most research studies can be classified into two categories: a) academic and demographic information, such as ACT/SAT scores, high school GPA, location, gender, ethnicity, first language, parents' education level, et cetera; and b) financial status like family income, tuition, financial aid, and scholarship offered. Though some research studies examined variables such as health and psychosocial condition, not all of these variables are readily available to all universities.

Chang (2008) presented a case study of data mining in enrollment management and used logistic regression to predict the student's probability of admission. This probability

is then utilized by the institution to form a referral pool of highly prospective students for direct marketing outreach.

Bruggink and Gambhir (1996) chose logistic regression to predict two stages in admission and enrollment. In the first stage, they predicted the probability of acceptance of each applicant. The results helped institutes to understand the features of potential incoming students such as geographical diversity, ethnicity, and academic performance. In the second stage, the enrollment probability of an admitted student was calculated. In this stage, they found that academically strong students are less likely to enroll.

Braunstein et al. (1999) analyzed the impact of financial and socioeconomic factors on enrollment. They found that, by using logistic regression, the probability of enrollment increased between 1.1% and 2.5% with every $1,000 increase of financial aid.

Kim (2004) applied logistic regression to see the impact of financial aid amounts of different racial groups. Their results have shown that overall, grants and combinations of loans and grants have a positive impact on enrollment; however, different impact patterns were found among individual racial groups.

Other studies in the individual enrollment prediction include Bayesian-based methods or decision-tree based analysis. For example, Thanh and Haddawy (2007) applied Naïve Bayes from Bayesian network to predict the probability of enrollment and developed an ensemble model to overcome highly skewed data to increase the accuracy of the prediction results.

Borah et al. (2011) proposed an application of decision tree models, specifically the C4.5 method with an attribute selection measure function to predict the enrollment of engineering students. In a similar binary prediction, Bailey (2006) used the classification

and regression tree (C&RT) algorithm to predict the graduation rate in a state's higher education system. The results provided information of Student-Right-to-Know disclosure and the leading factors of retention and graduation rates for decision makers.

The Critical Integration of Prediction and Optimization. Though there have been many papers on the effect of financial aid on student demand and college choices, and the solution to the problem is economically important to universities, neither the economic nor marketing literature particularly addresses how to integrate enrollment predictions with financial aid allocation together to optimize enrollment goals. These two problems are two critical components of one problem and neither should be omitted. On the one hand, prediction without following financial aid optimization would achieve little financial benefits; yet on the other hand, the optimization problem cannot be of any value if the prediction of response to financial aid from students with various social, financial, and academic factors are not available. In the next sections, the techniques of a three-phase approach to solve the optimal scholarship allocation problem are presented.

Predictive Models for Probabilities of Enrollment and of Graduation

Recall that the first phase of the proposed research is the predictive models for the probability of enrollment, length of studies, and the probability of graduation. The prediction models for the probabilities are based on some variant of logistic regression and presented in this chapter, and the prediction models for the length of studies are based on some variant of regression analysis and are presented in the next chapter.

3.1 Data Exploration and Visualization

The data used in this study were provided by the institutional research department at the university and consists of applications that span seven years from 2006 to 2013.

Application Counts. In these seven years, there are a total of 47,932 applications; 8,072 applications were not granted admission due to incomplete information or not meeting the admission requirements, and 35,331 were granted admission. Out of the 35,331 admitted

applications, 17,180 were enrolled, 18,151 were not; 44,994 (93.8%) of the applications are from instate, 2,971 are from out-of-state. As financial aid awards for out-of-state (including international) students are different from in-state students, these out-of-state applicants are removed from the analysis. Home-schooled applications (104) and applications to a satellite campus (1,796) are also removed from the analysis. This leaves a total of 35,331 applications, and the statistics of the number of enrolled (yes) and not enrolled (no) applications in each year are shown in Table 3.1.

	06-07	07-08	08-09	09-10	10-11	11-12	12-13
Enrolled	2,133	2,311	2,473	2,477	2,706	2,779	2,301
Not Enrolled	2,092	2,197	2,559	2,525	2,843	2,976	2,959

Table 3.1: The number of applications from 2007 to 2013

Variables. The variables associated with these applications can be classified into the following categories: academic performance, financial status, and personal information. The academic performance category includes the applicant's SAT/ACT scores, GPA, high school, high school report card, and high school percentile. The financial status category includes family income, expected family contribution (EFC), and Pell Grant award. The personal information category includes gender, ethnicity, distance from the institution, and the applicant's intended colleges and majors.

In the academic performance fields, the ACT score was normalized by translating it into a percentile. In the financial status fields, expected family contribution (EFC) measures an applicant's family's financial strength. The Federal Pell Grant is a need-based

grant for low-income students and is dependent on EFC. To better estimate an applicant's affordability, two more variables are introduced: total free money and out of pocket. The former defines all the awards that an applicant gets while the latter is the total out-of-pocket money an applicant pays.

Distribution of applications on categorical variables. The distributions of applications across selected categorical variables are presented below.

Gender: Among the applicants, 60% of them are female, and 40% are male.

Race: There are six categories of races reported: White (10,949 enrolled, 13,917 not enrolled), African American (2,916 enrolled, 4,400 not enrolled), Hispanic (355 enrolled, 461 not enrolled), and Asian (319 enrolled, 378 not enrolled). The rests are various others or not disclosed (673 enrolled, 959 not enrolled).

Distance Bin: There are six tiers, categorized by how far the student lives from the campus: Tier 1 (7,190 enrolled, 6,322 not enrolled), Tier 2 (2,183 enrolled, 2,909 not enrolled), Tier 3 (606 enrolled, 672 not enrolled), Tier 4 (2,146 enrolled, 4,446 not enrolled), Tier 5 (1,145 enrolled, 2,640 not enrolled), and Tier 6 (1,943 enrolled, 3,149 not enrolled)

Intending Colleges. Business (1,636 enrolled, 2,167 not enrolled), Science and Math (2,330 enrolled, 2,931 not enrolled), Liberal Arts (2,714 enrolled, 3976 not enrolled), Education (1,419 enrolled, 1,973 not enrolled), Engineering (2,081 enrolled, 2,347 not enrolled), Nursing (2,033 enrolled, 2,675 not enrolled), and University College (3,000 enrolled, 4,049 not enrolled).

Distribution of applications on continuous variables. The distributions of number of applications (n), mean, standard deviation (sd), median, minimum, and maximum of

applications, across selected continuous numeric variables, are presented in Table 3.2. The meanings of variables are self-explanatory and thus not elaborated upon.

Variable	n	mean	sd	median	min	max
GPA	35331	3.094	0.63	3.10	0.40	4.90
ACT	35331	21.27	4.39	21.00	8	36
HS PERCENTILE	35331	60.89	24.83	64.00	0	100
DISTANCE.NUM	35331	57.67	53.49	43.82	2.65	276.75
SCHOLARSHIP	35331	777.45	1339.30	0	0	8354
PELL GRANT	35331	1452.15	2228.13	0	0	5550
TOTAL FREE MONEY	35331	2229.61	2492.47	150	0	1390
OUT OF POCKET	35331	5455.20	2467.15	6278	-5935	8354

Table 3.2: Statistics of selected continuous variables related to applications

Academic performance wise, among the applications, the average student has a GPA of 3.09 and a composite ACT score of 21.27 (the average composite ACT score of the state is 22). The average high school percentile is 60.89%, and the average student lives 57.67 miles away.

Financially, among the applicants, the average awarded scholarship is $1452, and the average total free money, defined as Pell Grant plus scholarship contribution, is $2229, The average out of pocket spending is $5455. The sum of total free money and out of pocket is $7685, i.e., the average tuition cost across this period.

Distribution of matriculated students on numeric variables. The distributions across selected categorical variables related to enrolled applications are presented in Table 3.3.

Variable	n	mean	sd	median	min	max
GPA	17,180	3.08	0.61	3.10	1	4.90
ACT	17,180	21.08	4.22	21	10	35
HS.PERCENTILE	17,180	60.02	24.69	62	0	100
DISTANCE.NUM	17,180	48.80	48.32	30.77	2.65	270.24
SCHOLARSHIP	17,180	744.56	1,348.68	0	0	8,354
PELL.GRANT	17,180	1,750.12	2,336.59	0	0	5,550
TOTAL.FREE.MONEY	17,180	2,494.69	2,564.35	2,000	0	13,902
OUT.OF.POCKET	17,180	5,174.04	2,531.66	5,797	-5,683	8,354

Table 3.3: Statistics of selected continuous variables related to matriculated students

Academic performance wise, among the matriculated students, the average student has a GPA of 3.08 and a composite ACT score of 21.08. The average high school percentile is 60.02%, and the average student lives 48.32 miles away.

Financially, the average matriculated student received $745 in scholarships, and the total free money received was $2,495. Total out of pocket spending was $5,174. As before, the sum of total free money and out of pocket is the average tuition across the seven-year period.

To give a visual picture of the distributions, Figure 3.2 and 3.1 show the histogram

of selected continuous variables such as GPA and ACT for applicants and for matriculated students. Here “yes” represents enroll and “no” represents not enroll.

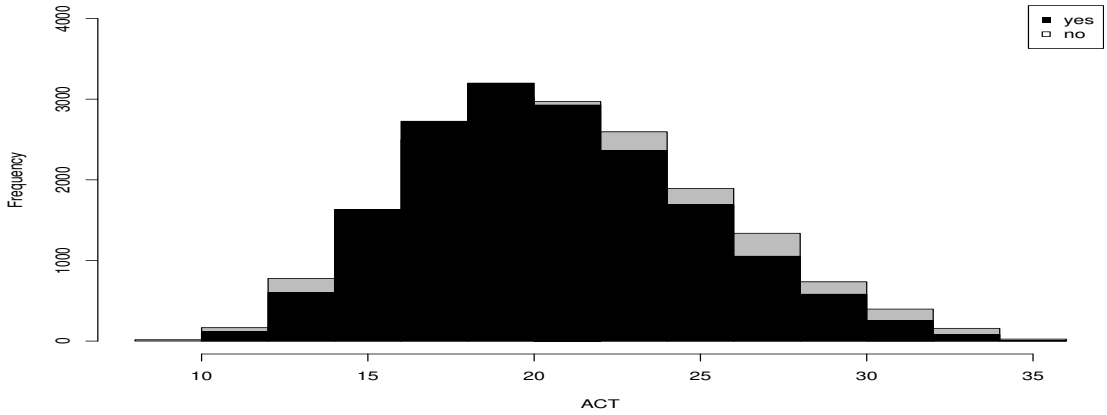

Figure 3.1: Histogram for ACT

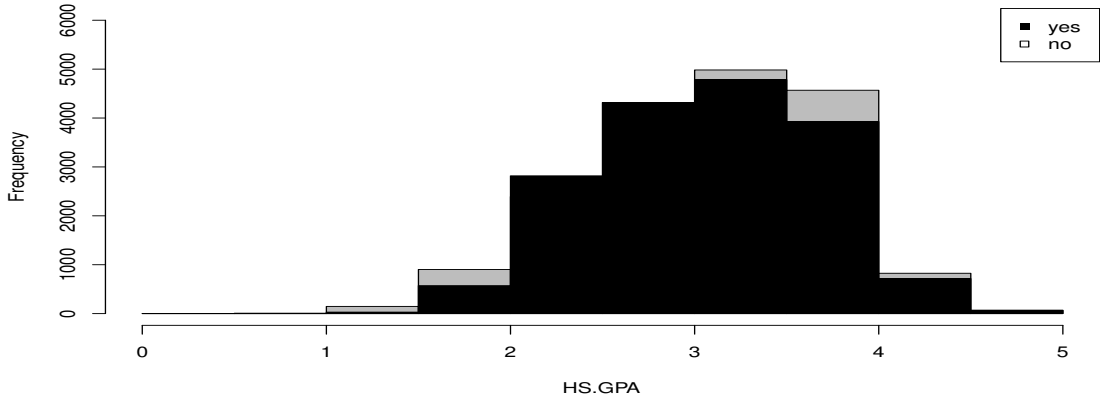

Figure 3.2: Histogram for High School GPA

The characteristics of matriculated students, such as GPA, ACT, etc, as can be seen, are similar to that of the applications. This is mainly because of the university’s open

admission policies and only a minor portion of incomplete applications are being filtered out.

Applicants Across Academic Measures. Table 3.4 shows the number of applicants with specific GPA and ACT scores. Here the number in the table represents the number of applications with the corresponding GPA (row) and ACT (column) scores. There seems to be a apparent correlation between GPA and ACT that will be discussed later.

GPA/ACT	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	Grand Total
1					1		2																					3
1.1									1																			1
1.2					1						1																	2
1.3				1				1		1																		3
1.4				1		2		1			2																	6
1.5				1	2	2	1	6	3		3							1										19
1.6				1	4	4	2	4	4	4	2	3	2	2	1	1												34
1.7	1			2	5	4	7	3	2	5	6	1	1	2	1	2	2	1										45
1.8					4	8	5	9	4	8	4	1	1	1														45
1.9		1	1	2	1	11	9	7	9	5	5	5	4	3	3		2											68
2				2	4	5	7	14	9	10	7	7	4	6	3	1	1				1							81
2.1	1			5	8	7	17	23	8	14	12	7	5	5	2		1	1	1			1						118
2.2			1	3	7	7	11	12	23	18	13	13	8	5	2	7	2	2	1	1	1							137
2.3			1	3	5	9	21	21	29	19	17	15	13	5	5	7	2	3	1	2								178
2.4				1	4	13	20	17	30	15	16	19	21	7	8	5	2	3	2			1						184
2.5				4	3	13	17	26	26	22	29	17	18	11	7	8	3	3	1	1	2		1					212
2.6			1	3	4	10	10	21	26	38	28	15	13	5	13	11	5	2	4		2	1			1			213
2.7			3		2	7	17	22	32	27	33	23	15	15	15	7	3	2	1	1		2						227
2.8				2	4	7	12	10	28	27	29	23	19	20	14	13	4	7	5	2	1				1			228
2.9			1	2	7	9	11	25	21	39	32	40	20	22	20	15	12	11	4	1								292
3				1	1	9	9	16	26	34	33	35	28	28	26	17	15	10	3	2	5		2			1		301
3.1				3	2	6	12	11	18	36	35	39	38	18	24	21	15	11	4	4	2	1					1	301
3.2						6	7	17	19	26	26	32	29	37	32	16	16	8	12	3	8	5	4	1				304
3.3			1		1	4	8	5	23	22	31	33	22	34	27	22	27	19	13	10	4	3	5		1	2		317
3.4		1			1		3	9	9	14	21	31	36	35	30	26	21	16	10	8	5	2	3					281
3.5							5	5	8	13	16	23	27	29	25	29	30	22	14	11	5	7	2		1			272
3.6							2	3	4	3	22	17	21	30	37	28	23	25	18	11	10	6	1	3	2	2		268
3.7						2	1		3	7	12	10	20	20	28	28	22	25	23	16	13	5	2	2		2		241
3.8							2			3	8	13	13	20	18	29	19	18	27	15	11	10	3	2	2			213
3.9							1	2	2	3	6	6	8	14	24	26	19	25	19	17	18	11	7	5	3	1		217
4								1			2	7	7	7	20	18	19	23	23	21	20	22	9	11	12	7	2	231
4.1									1		2	2	1	3	7	8	4	6	9	4	4	6	4	2	3	1		67
4.2										1		1	2	4	1	3	5	5	7	3	4	3	2	5	1	1		48
4.3														1	2	4	3	2	2	8	3	3	4	4	1	1		38
4.4																1	4	2	3	1	1	5	2	2	1	2		24
4.5												1		1	4		1	1		1	3		1	2	3		1	19
4.6															1			1	1	1		1		3	1	1		10
4.7													1					1	3	1			1	1	1			9
4.8																			1	1		1						3
Grand Total	2	2	9	37	71	145	219	291	368	414	453	439	397	390	400	353	282	256	212	146	123	96	53	43	34	21	4	5260

Table 3.4: Number of applicants vs GPA/ACT in 2012-2013

3.2 Logistic Regression For Enrollment & Graduation

3.2.1 Logistic Regression Methodology

Logistic regression is a popular method when predicting variables with dichotomous outcomes (yes and no) such as enroll (yes) and do not enroll (no). The output of a logistic regression model is the probability of the enrollment levels. Specifically, in this research, the probabilities to be sought are:

Enrollment probability: what is the probability of enrollment (yes) for an applicant with, say, a GPA of 2.5, an ACT of 28, given a $1,000 scholarship?

Graduation probability: what is the probability of graduation (yes) for the same applicant?

Let us denote $p^e(x)$ and $p^g(x)$ as the enrollment and graduation probability respectively. Because the probability $p(x)$ must fall in $[0, 1]$th, a regular linear regression is unsuitable as the regression function is not bounded. To resolve this issue, a ratio, called the odds of success and defined as $\frac{p(x)}{1-p(x)}$, is used to form a regression model and and a logistic function is defined as:

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{3.1}$$

The probability of enrollment or graduation can thus be rewritten as:

$$p(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_i x_i}} = \frac{1}{1 + e^{-(\beta_0+\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k)}} \tag{3.2}$$

Here, β_0 is the intercept and $\beta_1 \ldots \beta_i$ are parameters for corresponding variable x_i. Here x_i could be either continuous variables such as GPA or ACT, or categorical variables such as gender, ethnicity, or region. Parameters β_0 and β_i are estimated by the maximum likelihood estimation.

$$\ell(\beta_0, \ldots \beta_i) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_i'=0} (1 - p(x_{i'})) \tag{3.3}$$

For details on the derivation of the parameter estimations, please see Hosmer et al. (2013).

3.2.2 Collinearity and Variable Selection

The baseline model is a general linear logistic regression model. Before the model is presented, collinearity among variables and variable selections are first explored.

Collinearity. Collinearity refers to the fact that two or more predictor variables within the regression models are highly correlated. Collinearity makes the estimate highly unstable, i.e., coefficient estimates are very sensitive to small changes due the model or data. As a result, it is hard to interpret which variables are contributing to the model and to identify how exactly each variable is contributing to the model (Belsley et al., 2005; Habshah et al., 2010). For this study, the correlation matrix, a common method to examine collinearity, of selected continuous variables is presented in Table 3.5.

	GPA	ACT	HS. PERCENTILE	SCHOLAR-SHIP	PELL. GRANT
GPA	1	0.596	**0.884**	0.612	-0.190
ACT	0.596	1	0.469	0.586	-0.275
HS.PERCENTILE	**0.884**	0.469	1	0.594	-0.060
SCHOLARSHIP	0.612	0.586	0.594	1	0.025
PELL.GRANT	-0.190	-0.275	-0.060	0.025	1

Table 3.5: Pearson correlation matrix of all numeric variables

As can be seen from Table 3.5, according to (Hinkle et al., 2003), there exists a) a high correlation between GPA and HS.PERCENTILE, b) a high correction between academic performance measures such as GPA and ACT, c) a high correlation between scholarship and academic performance measures such as GPA and ACT, and d) a negative correlation between Pell Grant and academic performance (such as GPA and ACT). Though some of the correlations may not be surprising, the last one seems to suggest that family income could impact a student's academic performance.

Collinearity does not reduce the fitness of the overall model, but it can lead to the erroneous prediction of power of individual predictor, which cause inaccurate interpretation of the model.

Variable Selection. In regression analysis with multiple variables, strategies have been designed to select the best variables in the model. Stepwise selection is one of the most

popular procedures and is adopted in this study (Konishi and Kitagawa, 2008).

The stepwise variable selection includes backward elimination, forward selection, and bidirectional elimination. A backward stepwise regression, for example, works as follows: a) start with the full model with all predictors; b) subtracting one predictor at a time, removing the predictor if doing so improves accuracy; and c) iterate until no further improvement.

To determine the best models in the process, several criteria have been proposed. For logistic regression, the most common ones are BIC, AIC, and AICc. In this paper, the AIC is adopted due to its popularity. There is always a trade-off between model simplicity and accuracy, and the AIC methods determine whether additional variables are justified. For details, please see (Wagenmakers and Farrell, 2004).

In an enrollment model, say with n predictor variables, to get the optimal model, a brute-force approach would have to evaluate 2^n possible models. The stepwise variable selection evaluates only a limited number of models and presents a heuristic solution to the problem in a fraction of the time.

3.2.3 Logistic Regression Models on Training Data

Logistic Regression Model for Enrollment Decision.

The data from academic years 2008-2009 to 2011-2012 is used as training data and the data from the year 2012-2013 as testing data. The training data has 30,083 entries and the testing data has 5,260 entries.

The summary statistics for major variables in the logistic regression selected by the

backward stepwise selection for enrollment are shown in Table 3.6. In the table, the "estimate" column reports the odds ratio obtained from logistic regression, and an entry with value larger (smaller) than 1 represents that as the corresponding value increases, the odds would increase (decrease). For other fields, please see (DesJardins et al., 2006; Long and Freese, 2006).

	Estimate	Std. Error	z value	Pr(>|z|)	Signif	2.50%	97.50%
(Intercept)	2.5934	0.1476	6.4600	0.0000	***	1.9423	3.4639
GPA	1.3405	0.0463	6.3300	0.0000	***	1.2242	1.468
ACT	0.9630	0.0038	-9.8200	0.0000	***	0.9558	0.9703
HS.PERCENTILE	0.9906	0.0011	-8.8100	0.0000	***	0.9885	0.9927
DISTANCE.BIN02	0.6498	0.0420	-10.2600	0.0000	***	0.5983	0.7054
DISTANCE.BIN03	0.4996	0.0737	-9.4100	0.0000	***	0.4323	0.5772
DISTANCE.BIN04	0.5819	0.0826	-6.5600	0.0000	***	0.4949	0.6841
DISTANCE.BIN05	0.4346	0.0843	-9.8900	0.0000	***	0.3683	0.5126
DISTANCE.BIN06	0.3029	0.1124	-10.6200	0.0000	***	0.243	0.3776
PELL.GRANT	1.0002	0.0000	27.5700	0.0000	***	1.0001	1.0002
ETHNICITYAfricanAmerican	1.0272	0.0870	0.3100	0.7581		0.8661	1.2182
ETHNICITYHispanic	1.1814	0.1085	1.5400	0.1245		0.955	1.4614
ETHNICITYOthers	0.4939	0.1265	-5.5800	0.0000	***	0.3848	0.632
ETHNICITYWhite	1.3417	0.0816	3.6000	0.0003	***	1.1433	1.5746
UNEMPLOYMENT.INDEX	0.9655	0.0059	-5.9100	0.0000	***	0.9544	0.9768
SCHOLARSHIP_PER	1.4894	0.0933	4.2700	0.0000	***	1.2403	1.7883

$^{***}p = 0.000, ^{**}p < 0.001, ^{*}p < 0.01, ^{\cdot}p < 0.05$

Table 3.6: Summary statistics of logistic regression for enrollment model

In interpreting these results, however, care should be taken because of the existence of collinearity. At first glance, it seems puzzling that applicants with lower ACT are less likely to enroll, that applicants with higher GPA are more likely to enroll, and that HS.PERCENTILE does not impact enrollment very much. This is likely because of the fact that HS.PERCENTILE, ACT, and GPA are highly correlated with each other, so their effects are masked in the model.

If an applicant is underrepresented, they are less likely to enroll. Distance bins except

Bin 1 (closest) are all less inclined to enroll. Scholarship (SCHOLARSHIP_PER) has an impact on the enrollment – the more scholarship that one receives, the more likely one is to enroll. The unemployment index (0.9847), representing economy, has little positive impact on the enrollment.

Logistic Regression Model for Graduation.

Table 3.7 shows the summary statistics of major variables in the logistic regression for the graduation model. The table is similarly organized to the logistic regression for enrollment models. It seems to suggest that GPA and scholarship both have significant impact on graduation. The higher the GPA and the higher the scholarship a student is awarded, the more likely that the student will graduate.

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	Significant	5%	95%
(Intercept)	2.6008	0.1576	3.1800	0.0015	**	1.2737	2.1390
GPA	2.6872	0.0505	7.9640	0.0000	***	1.3757	1.6242
ACT	2.3540	0.0042	-7.5840	0.0000	***	0.9622	0.9755
HS.PERCENTILE	3.4288	0.0012	-8.1700	0.0000	***	0.9886	0.9924
DISTANCE.BIN02	1.6403	0.0446	-8.8600	<2e-16	***	0.6256	0.7246
DISTANCE.BIN03	1.4453	0.0794	-7.4110	0.0000	***	0.4870	0.6325
DISTANCE.BIN04	1.2751	0.0889	-5.3070	0.0000	***	0.5391	0.7221
DISTANCE.BIN05	2.1893	0.0911	-7.4790	0.0000	***	0.4356	0.5877
DISTANCE.BIN06	2.4776	0.1223	-8.0080	0.0000	***	0.3073	0.4594
HS.COUNTY.TIERTier2	1.6972	0.0534	-3.9420	0.0001	***	0.7418	0.8844
HS.COUNTY.TIERTier3	1.8727	0.1002	0.2980	0.7657		0.8738	1.2149
HS.COUNTY.TIERTier4	2.0888	0.0873	-5.9900	0.0000	***	0.5135	0.6844
HS.COUNTY.TIERTier5	2.7186	0.1101	-3.4020	0.0007	***	0.5736	0.8240
HS.COUNTY.TIERTier6	2.3776	0.0859	-3.0990	0.0019	**	0.6652	0.8826
PELL.GRANT	1.1828	0.0000	22.9390	0.0000	***	1.0001	1.0002
ETHNICITYBlackOrAfricanAmerican	2.2082	0.0925	1.1540	0.2483		0.9558	1.2958
ETHNICITYHispanic	4.2355	0.1170	1.8430	0.0654	.	1.0234	1.5040
ETHNICITYWhite	3.4567	0.0864	3.0690	0.0021	**	1.1312	1.5032
UNEMPLOYMENT.INDEX	1.0000	0.0060	-9.7690	<2e-16	***	0.9336	0.9522
SCHOLARSHIP_PER	1.0000	0.1010	5.1330	0.0000	***	1.4221	1.9823

$^{***}p = 0.000, ^{**}p < 0.001, ^{*}p < 0.01, ^{\cdot}p < 0.05$

Table 3.7: Summary statistics of logistic regression for graduation model

3.2.4 Logistic Regression Tree Models on Training Data

Logistic Regression Tree Methodology

A logistic regression tree model extends the baseline logistic regression model and uses a *divide and conquer* strategy to divide the data into many subsets so that a logistic regression model fits the data in each subset. And more logistic regression was generated recursively using the subset data and finally results in the partitions of a binary decision tree (Harrell, 2013).

The logistic tree with unbiased selection (LOTUS) algorithm (Chan and Loh, 2004) is used in this research for the automation of a logistic regression tree. LOTUS allows nonlinear variables to be modeled and outperforms the standard stepwise logistic regression (Chen et al., 2013; Yamashita et al., 2013).

Logistic Regression Tree Models

Two logistic regression trees are constructed: one for enrollment, one for graduation. In the construction of these tree models, the variables used in these two tree models and their role in the tree models are listed in the table below. Here, "D" represents the dependent variable, "S" is used as a numerical variable to split the tree, "C" represents categorical variables to split the tree, "X" represents variables to ignore, and "F" represents the decision variables used in the logistic regression function at the tree node.

Column	Enrollment		Graduation	
	Name	Type	Name	Type
1	Enrolled	D	Enrolled	X
2	GPA	S	GPA	S
3	Tier	C	Tier	C
4	Raider	C	Raider	C
5	ACT	S	ACT	S
6	Underrepresented	C	Underrepresented	C
7	Gender	C	Gender	C
8	Ethnicity	C	Ethnicity	C
9	Scholarship($)	F	Scholarship($)	F
10	Scholarship(%)	F	Scholarship(%)	F
11	Graduate	X	Graduate	D

Table 3.8: Variables used in the logistic regression tree models

3.2.5 Prediction Accuracy on Test Data

A. Logistic Regression and Logistic Regression Tree Models

The logistic regression model and logistic regression tree models derived from the test data are applied to the test data to verify the accuracy of the models. The cut-off value for

enrollment is set to 0.5. So if $p(x) \geq 0.5$, then enroll; otherwise, not enroll.

For the logistic regression model for enrollment prediction, the accuracy of the logistic regression model is 0.619 while the AUC is 0.658, as seen Table 3.9. For the logistic regression model for graduation prediction, the accuracy of the logistic regression model is 0.74 while the AUC is 0.79, and seen in Table 3.10.

Model	AUC	Accuracy	Enrolled	True Positive/ Negative Rate	False Positive/ Negative Rate
Logistic Regression	0.658	0.619 *	Yes	0.646	0.414
			No	0.586	0.354
Logistic Regression Tree	0.618	0.62 *	Yes	0.65	0.414
			No	0.586	0.35

Table 3.9: Enrollment prediction from logistic regression and logistic regression tree

Model	AUC	Accuracy	Graduated	True Positive/ Negative Rate	False Positive/ Negative Rate
Logistic Regression	0.79	0.74 *	Yes	0.8	0.41
			No	0.59	0.2
Logistic Regression Tree	0.76	0.739 *	Yes	0.88	0.57
			No	0.43	0.12

Table 3.10: Graduation prediction from logistic regression and logistic regression tree

B. Support Vector Machines and Neural Networks

Preliminary experience with other logistic based models such as support vector machines and neural networks suggest similar accuracy results. The accuracy of prediction of these models are shown in the Table 3.11 and 3.12.

Model	AUC	Accuracy	Enrolled	True Positive/ Negative Rate	False Positive/ Negative Rate
Support Vector Machine	0.612	0.608 *	Yes	0.531	0.306
			No	0.694	0.469
Neural Network	0.611	0.616 *	Yes	0.647	0.425
			No	0.575	0.353

Table 3.11: Accuracy of enrollment prediction from support vector machine and neural networks

Model	AUC	Accuracy	Graduated	True Positive/ Negative Rate	False Positive/ Negative Rate
Support Vector Machine	0.66	0.72 *	Yes	0.77	0.230
			No	0.57	0.43
Neural Network	0.68	0.737 *	Yes	0.78	0.22
			No	0.60	0.40

Table 3.12: Accuracy of graduation prediction from support vector machine and neural network

Due to space limitations, details of these models are not represented. To further develop these models to increase the prediction accuracy is the focus of future study.

3.3 Answer to Enrollment & Graduation Probabilities

The predictions of enrollment probabilities for six applications with different levels of scholarship awards are listed in Table 3.13.

GPA 2.9, ACT 19											
	Student	0	1000	2000	3000	4000	5000	6000	7000	8000	10000
1	2.9-Tier1-19-White	59.55	64.63	69.39	73.77	77.73	81.24	84.31	86.96	89.22	91.72
2	2.9-Tier5-19-White	36.96	40.20	43.53	46.92	50.34	53.75	57.13	60.45	63.67	69.74
GPA 3.3, ACT 25											
3	3.3-Tier1-25-Hispanic	23.80	27.44	31.42	35.69	40.20	44.88	49.65	54.43	59.13	67.97
4	3.3-Tier1-25-White	55.60	59.32	62.94	66.42	69.72	72.84	75.75	78.43	80.90	85.17
GPA 3.8, ACT 28											
5	3.8-Tier1-28-White	42.29	46.05	49.85	53.65	57.41	61.08	64.63	68.03	71.25	77.07
6	3.8-Tier4-28-White	20.54	22.87	25.37	28.05	30.89	33.89	37.02	40.26	43.60	50.41

Table 3.13: Prediction of enrollment under different levels of scholarships

Here, the concatenated string under “student” represents the characteristic of the student. For example, student “2.9-Tier1-19-White” represents an application from a student with a high school GPA of 2.9, ACT score of 19, lives in Tier 1 region, and is of the white.

The number represents the probability of enrollment given the corresponding scholarship award, which spends from $0 to $10,000.

Observations. Several interesting observations can be seen from these predictions, a) as GPA increases, the probability of enrollment decreases; b) local students (Tier 1) have a larger probability of enrollment than distance students (other tiers); c) as financial aid increases, the probability of enrollment increases, yet increase in probability with respect to scholarship is different among different student groups.

- Students 1 and 2: both students have the same GPA, ACT, and ethnicity, but student 1, who lives in Tier 1, has a much higher probability of enrollment than student 2, who lives in Tier 5.
- Students 1 and 5: both students live in the same region and are of the same ethnicity, but student 1 (GPA of 2.9, ACT of 19) has a higher probability of enrollment than student 2 (GPA of 3.8 , ACT of 28).
- Students 3 and 4: both students live in the same region and have the same GPA and ACT scores, but student 3 has a higher probability of enrollment than student 4 due to different ethnic backgrounds.

In all these cases, the probability of enrollment increases with the increase of scholarship awards. Though these observations are not surprising, accurate quantitative prediction of these probabilities is essential to the allocation of scholarship.

However, it bears mentioning that the prediction of these probabilities alone has not yet solved the fundamental problems for the enrollment management team of any higher

institution; the optimal allocation of the scholarship to optimize an institution's revenue, and the formation of concrete policies and action plans, needs to be solved. This is the second part of the research and is addressed in the following sections.

Prediction Models on the Number of Years of Study

Recall that the net tuition of a student is the difference between tuition and scholarship, and the scholarship is renewable. This chapter studies the models to address the question on the number of years of study once a student enrolls in the university.

4.1 Difference From a Retention Study

It bears mentioning that the prediction on the number of years of study is is different from retention studies in most literature reviews. In retention studies, the focus is what factors influence the retention, and by doing analysis, early stage hazards can be identified to improve retention rate.

Retention and dropout is affected by the interaction of students' pre-enrollment characteristics (academic performance, finance, ethnicity, etc.) and academic experience (peer group interactions, interaction with faculty, etc.) (Tinto, 1975, 1982; Terenzini et al.,

1981). Studies found that first-year students are the most vulnerable to drop out. Freshmen face drastic changes not only in the form of the academic challenge but also all kinds of social challenges. Therefore retention of first-year students is mostly studied (Permzadian and Credé, 2016; Kovačić, 2010; Horstmanshof and Zimitat, 2007; Noble et al., 2007). Classification prediction methods such as logistic regression, decision tree, random forest, and neural networks were mostly used (Dekker et al., 2009; Adam and Gaither, 2005; Quadri and Kalyankar, 2010; Yu et al., 2010; Herzog, 2006; Lin et al., 2009; Zhang et al., 2010; Herzog, 2006).

In a retention study, besides pre-collegiate information, variables such as campus experience (on or off campus living, average class size, etc.), college academic performance (credit hours taken, grades, etc.), and ongoing financial need are critical. In addition, health and psychosocial variables such as smoking, drinking, health-related quality of life and social support, were found significantly related to academic achievement and retention (DeBerard et al., 2004; Maney, 1990; Musgrave-Marquart et al., 1997; Cutrona et al., 1994).

This study, however, is mainly focused on the use of pre-collegiate information such as demographics, high school academic performance, and financial background to predict the number of years study.

4.2 Methods and Results

4.2.1 Training and Testing Data

In this study, only records of enrolled students during the 2006, 2007, and 2008 academic years were utilized, as the the latest data obtained is Fall semester 2012 and the cutoff value for the study is graduation within 5 years. Among the enrolled students, 43% were male, 57% were female; 73.9% Caucasian, African American 17.8%, 2.1% Asian, 2% Hispanic, 4.2% Others.

The training data of this study is from the years 2006-2007 and 2007-2008 when 3,999 students were enrolled. The testing data is from the year 2008-2009 when 2,355 students were enrolled. Variable selection for the number of years prediction is similar to the enrollment and graduation prediction. Again, unlike other studies (Lin et al., 2009; DeBerard et al., 2004; Dekker et al., 2009), which tracked the after enrollment data such as semester GPA and social activities, this research only used pre-collegiate variables due to data collection limit.

4.2.2 Prediction Models

Various data mining models such as linear regression, support vector machine regression, random forest, CART, and stochastic gradient boosting are used to predict the number of years of study and provide ranks on the importance of the variable and are presented below.

Generalized Linear Model (GLM) is a generalization of the linear regression model with

a normally distributed dependent variable and Gaussian error. GLM broadens the distribution of the dependent variable to another family such as exponential or binomial. GLM can analyze interactions of variables, including mixtures of categorical and continuous variables.

Support Vector Machines (SVMs) are supervised classifiers used for classification, regression and outliers detection. Given labeled data, the algorithm outputs an optimal hyperplane or set of hyperplanes, which separate the data into different classes. It also can be used for regression purposes.

Following is a simple two-dimensional example where the classes are linearly separable. We have labeled example (x_1, y_1),...,(x_n, y_n) with label $y_i = 1$ for inputs x_i in class 0 and $y_i = -1$ for inputs x_i in class 1. The classification boundary or hyperplane is defined as $w^T x + b = 0$, where w is the weight vector and b is the bias. The hyperplane can be represented by different scales of (w, b). The optimal hyperplane is defined as $|w^T x + b| = 1$, where x is the data points closest to the hyperplane; for instance, the negative classification boundary is $w^T x + b = -1$ and positive classification boundary is $w^T x + b = 1$. As a result, the distance between data point x and the hyperplane (w, b) is $\frac{|w^T+b|}{||w||} = \frac{1}{||w||}$, so the total distance to positive and to negative class, defined as the margin M, is $\frac{2}{||w||}$. The goal is to maximize margin M to separate the data points valued 1 from those having -1, so we have to minimize the w^T. The final problem of linear SVM is to optimize:

$$\text{minimize } \frac{2}{||w||^2}, \text{ subject to } y_i(w^T x + b) \geq 1 \; \forall i \tag{4.1}$$

where y_i is either positive class (1) or negative class (-1).

Support Vector Machine Regression (SVM regression) is a sub-category of SVM to solve regression problems. SVM allows the use linear regression in the high-space by using ϵ-insensitive loss and aims to reduce model complexity by minimizing $||w||^2$. By introducing slack variables ξ, the model is able to measure the error of training data outside the ϵ-insensitive zone. Thus, SVM regression is all about solving the following minimization problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||w||^2 + C\sum_{n=1}^{n}(\xi_i + \xi_i^+) \\ \text{subject to} \quad & y_i - f(w, x_i) \leq \epsilon + \xi^+, \\ & f(w, x_i) - y_i \leq \epsilon + \xi, \\ & \xi^i, \xi^+ \geq 0, i = 1, \ldots, n \end{aligned} \tag{4.2}$$

where x_i is the training data and y_i is target variable.

Decision Trees are commonly used to build classification and regression models in the form of tree-like shapes. Decision tree methods separate data into groups to grow branches. A decision tree contains root nodes, terminal nodes, and internal nodes. The root node is the topmost node. Leaf nodes contain the final decision values of the dependent variable. Internal nodes represent the values of attributes. The algorithms used to build a decision tree include ID3, C4.5, C5.0, CHAID, CRUISE, etc. To split the tree, measurement algorithms like Gini index or information gain is used. For regression specifically, minimized

residual sum of squares (SSE):

$$SSE = \sum_{i \in S_1} (y_i - \overline{y_1}) + \sum_{i \in S_2} (y_i - \overline{y_2}) \tag{4.3}$$

are used to split the value of an attribute. $\overline{y_1}$ and $\overline{y_2}$ are the average values of the dependent variables in group S_1 and S_2.

Stochastic Gradient Boosting is a sub-category of boosting methods which convert weak learners to strong learners. A boosting model works in the following way: starting with a base machine learning algorithm with a different distribution, a second model is generated based on correcting errors from the first model. A decision tree with a fixed sized is typically used as the weak learner in the gradient boosting. The process iterates until the limit of the base algorithm is reached, or a certain accuracy is achieved. The estimate of response variables becomes consecutively more accurate between iterations. Stochastic gradient boosting for regression problems uses square error as a loss function. At each iteration, a sample of data was randomly selected from the full training data without replacement. The weaker learner then is built on the sample data instead of the full training data (Friedman, 2002).

Model Metrics. To evaluate the performance of the model, the Root Mean Square Error (RMSE) and the mean absolute error (MAE) are commonly used. RMSE is the square root of the average square of the difference between our predicted and actual values. RMSE is

in the same units as the predicted value.

$$RMSE : \sqrt{\frac{1}{n}\sum_{n=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.4}$$

MAE is used to measure the difference of forecasts to the real outcomes.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.5}$$

Model Implementation. For SVM regression implementation, linear kernel in R's e1071 package (Meyer et al., 2017) using default settings. For GLM, base function in R base package was used. For stochastic gradient boosting, gbm package in R (Ridgeway, 2017) was used.

4.2.3 Experiment Results

Two sets of experiments were conducted to evaluate the model's performance. The first experiment is 10-fold cross-validation. Data used in this experiment is from the year 2006-2007 and 2007-2008. 80% of the data was randomly chosen to build the model, and 20% percent of the data was used to validate the model in each fold. The final evaluation model is the average metric of all the folds. The second experiment is using data from 2008-2009 to predict the unseen students. The results of the prediction models are as shown below:

	10-Fold Cross Validation		Test Data	
Model	RMSE	MAE	RMSE	MAE
GLM	1.40	1.2	1.53	1.26
SVM(Linear Kernel)	1.44	1.20	1.62	1.32
Decision Tree	1.43	1.24	1.43	1.23
Stochastic Gradient Boosting	1.40	1.19	1.40	1.19

Table 4.1: Model results of number of years of study

The results show that the tree-based methods (decision trees and stochastic gradient boosting) yield slightly better results with lower RMSE and MAE for both 10-fold cross validation and validation based on test data.

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	Significant	5%	95%
(Intercept)	-3.0276	1.3632	-2.2	0.026	*	0.0033	0.7006
GPA	0.73784	0.0510	14.5	<2e-16	***	1.8922	2.3115
ETHNICITYBlackOrAfricanAmerican	0.12447	0.1744	0.7	0.475		0.8046	1.5941
ETHNICITYHawaiian	0.25884	1.0005	0.3	0.796		0.1822	9.2056
ETHNICITYHispanic	0.50478	0.2318	2.2	0.030	*	1.0517	2.6094
ETHNICITYIndianAlaskan	0.01794	0.4061	0.0	0.965		0.4431	2.1771
ETHNICITYTwoMore	1.05399	0.2341	4.5	7e-06	***	1.8132	4.5396
ETHNICITYUnknown	0.79924	0.2364	-3.4	7e-04	***	0.2828	0.7148
ETHNICITYWhite	0.03123	0.1623	-0.2	0.847		0.7050	1.3323
HS.COUNTY.TIERTier2	0.23719	0.0701	-3.4	7e-04	***	0.6874	0.9051
HS.COUNTY.TIERTier3	0.08538	0.1217	-0.7	0.483		0.7232	1.1656
HS.COUNTY.TIERTier4	0.24058	0.0820	-2.9	0.003	**	0.6693	0.9233
HS.COUNTY.TIERTier5	0.22394	0.0923	-2.4	0.015	*	0.6670	0.9579
HS.COUNTY.TIERTier6	0.15317	0.0750	-2.0	0.041	*	0.7406	0.9938
APP.COLLEGEED	0.27311	0.1005	-2.7	0.007	**	0.6249	0.9267
APP.COLLEGEEG	0.12430	0.0956	-1.3	0.194		0.7320	1.0652
APP.COLLEGELA	0.12102	0.0883	-1.4	0.171		0.7450	1.0535
APP.COLLEGEN	0.33059	0.0972	-3.4	7e-04	***	0.5938	0.8693
APP.COLLEGESM	0.22383	0.0952	-2.4	0.019	*	0.6633	0.9635
APP.COLLEGEUC	0.17456	0.0857	-2.0	0.042	*	0.7099	0.9935
OUT.OF.POCKET	0.00004	0.0000	3.5	5e-04	***	1.0000	1.0000
SCHOLARSHIP.PER	0.62733	0.2033	3.1	0.002	**	1.2570	2.7896
UNEMPLOYMENT.INDEX	0.58469	0.2416	2.4	0.016	*	1.1175	2.8814

$^{***}p = 0.000$, $^{**}p < 0.001$, $^{*}p < 0.01$, $^{\cdot}p < 0.05$

Table 4.2: Summary statistics of linear regression for number of years of study

Variable Importance. Finally, Table 4.3 presents the importance of variables in the prediction of number of years from a boosting model.

Variable	Relative Influence
GPA	47.91771887
HS.PERCENTILE	12.9595027
ACT	7.050451247
PELL.GRANT	4.743360342
SCHOLARSHIP_PER	4.033307548
UNEMPLOYMENT.INDEX	3.202687597
OUT.OF.POCKET	1.770385883
DISTANCE.BIN04	1.251210899
DISTANCE.BIN05	0.893538412
DISTANCE.BIN02	0.79644945
DISTANCE.BIN03	0.788177176
ETHNICITYWhite	0.542182288

Table 4.3: Relative influence of variables in gradient boosting model

Mathematical Models for The Optimal Financial Aid Allocation Problem

The study of students' responses to scholarship awards, though it has provided many insights into the behavior of students in responding to awards, has not addressed the allocation of limited financial aid to students fundamentally.

For example, it is easy to see that local students require less money while students in far-away regions may need more money, but it is still puzzling as: a) should we allocate the money to local students as they are our bread and butter students and require less money or b) should we allocate the money to far-away students as local students will come anyway? The solution to these problems requires the solution of an optimization problem to allocate the financial aid optimally, which is addressed in this chapter.

5.1 The Financial Aid Optimization Model

Given a set of applicants and their probabilities of enrollment and graduation with respect to different levels of financial aid, the optimization problem to be solved is to determine the financial aid to each applicant to maximize the revenue. This is referred as the financial aid allocation problem.

In the development of the model, the following notations are used:

Sets

I	set of applicants, indexed by i and j
M	set of different levels of financial awards, indexed by m
	$m \in M = \{0, 1000, 2000, \ldots, 8000\}$

Parameters

p^e_{im}	probability of enrollment for applicant i, if given award m
p^g_{im}	probability of graduation for applicant i, if given award m
N_{im}	expected number of years student i stays at the institution, if given award$m$
$d(i, j)$	1 if applicant i dominates applicant j; 0 otherwise.
B	total budget for financial aid
A_m	monetary value of award m
T_i	tuition paid by applicant i
SSI_i	government compensation for applicant i when he/she graduates

Variables

x_{im}	whether a financial award m is allocated to applicant i or not

Objective

$$\max \quad \sum_{i \in I} \sum_{m \in M} x_{im} \cdot p^e_{im} \cdot (T_i - A_m) \cdot N_{im} + \sum_{i \in I} \sum_{m \in M} x_{im} \cdot p^e_{im} \cdot p^g_{im} \cdot SSI_i \tag{5.1}$$

Subject to

$$\sum_{m \in M} x_{im} = 1 \qquad \forall i \in I \tag{5.2}$$

$$\sum_{i \in I} \sum_{m \in M} x_{im} \cdot p^e_{im} \cdot A_m \leq B \tag{5.3}$$

$$\sum_{m \in M} x_{im} \cdot A_m \geq \sum_{m \in M} x_{jm} \cdot A_m \qquad \forall (i,j) | d(i,j) = 1 \tag{5.4}$$

The objective (5.1) is to maximize the total revenue. The first term represents the total tuition income from matriculated students, i.e., the tuition minus the award for each student, times the number of years of study; the second term represents the state compensation once the student graduates.

Constraint 5.2 states that each applicant is given one award (zero is an award with no monetary value). Constraint 5.3 states that the total financial aid allocated cannot exceed the total budget B. Constraint 5.4 states that if applicant i dominates applicant j, then applicant i should be allocated a higher level award than applicant j.

5.2 Model Size Reduction and Dominance Matrix

5.2.1 The Size of Pair-wise Dominance Constraints

The size of the above model could be very large because of the number of pair-wise dominance relationships that form constraint 5.4. For example, the university under study typically has more than 5,500 applicants each year; for each applicant i and j ($i \neq j$), there will be $(5,500 \times 5,500)/2$ or more than 15 million constraints. Initial experiments with state-of-the-art commercial solvers were unsuccessful due to running out of memory.

However, if an applicant i dominates applicant k, and applicant k dominates applicant j, then it is only necessary to explicitly include domination constraints for applicants i and k, as well as k and j, but not necessarily for i and k. The dominance constraint i and k is redundant as it is implicitly expressed in the other constraints. In view of this, to reduce the size of the model, an efficient algorithm has been developed to find the domination matrix of minimum cardinality without redundant dominance.

5.2.2 Full Dominance Matrix

The full dominance matrix between any applicant is defined first below. **Full (Direct) Dominance Matrix:** Let D^f be an n by n matrix, where each element $d_{i,j}$ represents whether applicant i dominates applicant j or not. For simplicity, dominance is only related to academic performance, for example:

$$d_{i,j} = \begin{cases} 1 & \text{if } GPA_i \geq GPA_j \text{ and } ACT_i \geq ACT_j \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

Example: Table 5.1 presents the ACT and GPA scores of six applicants. Based on the above dominance definition, among these applicants, applicant 2 dominates 1, 4 and 5; applicant 3 dominates 1, 2, and 4; applicant 6 dominates applicants 1, 2, 4 and 5.

Applicant	GPA	ACT
1	2.9	18
2	3.7	21
3	3.8	30
4	2.7	21
5	3.3	17
6	3.9	27

Table 5.1: An example six students and their GPA and ACT scores

These pairwise dominance relationships can be represented by graph and matrix forms shown in Figure 5.2. Here, an arc between applicant i and j in the graph represents the dominance of applicant i over j, as is the entry of 1 in cell (i, j) in the matrix form.

In this example, applicant 2 dominates applicant 1 and applicant 3 dominates appli-

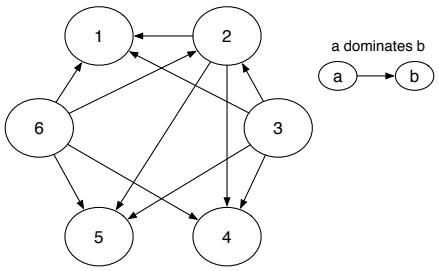

Figure 5.1: Full dominance relationships in graph form

	1	2	3	4	5	6
1	0	0	0	0	0	0
2	1	0	0	1	1	0
3	1	1	0	1	1	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	1	1	0	1	1	0

Table 5.2: Full dominance relationships in matrix form

cant 2, so the dominance between applicant 3 and applicant 1 is redundant and can be eliminated.

5.2.3 Redundant Dominance Matrix

Graphically, a redundant relationship from node i to node j states that there exists at least one two-step path (not a direct path from node i to node j) with one intermediate node, say from node i to node k, and then from node k to node j (where k can be any intermediate node). The number of two-step paths from node i to node j with one intermediate node can be easily calculated by

$$\sum_{k} d(i,k) \cdot d(k,j)$$

i.e., if there exists a redundant relationship, the inner product of the two corresponding vectors should be greater or equal to 1, and 0 otherwise.

The redundant relationship can thus be represented in a matrix, denoted as D^2, as follows.

$$D^2 = D^f \cdot D^f$$

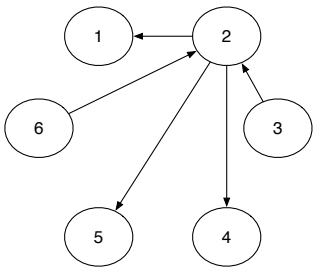

Figure 5.2: Redundant dominance relationships in graph form

	1	2	3	4	5	6
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	1	0	0	1	1	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	1	1	0	1	1	0

Table 5.3: Redundant dominance relationships in matrix

Here D^f is the original direct dominance matrix and each entry in D^2 represents the number of two-step paths between a pair of applicants.

Example: Applicant 2 dominates applicant 1, and applicant 3 dominates both applicants 1 and 2, therefore the entry $d_{21} = d_{31} = d_{32} = 1$. The relationship between applicant 3 and the other applicants k is: $d_{3k} = (1, 1, 0, 1, 1, 0)$, and applicant 1 and the other applicants is $d_{k1} = (0, 1, 1, 0, 0, 1)^T$. Furthermore $\sum_k d_{3j} \cdot d_{k1} = d_{31} = 1$. This represents that the relationship between applicants 1 and 3 is redundant.

Finally, the elements of a **redundant matrix**, denoted as D_r, can be defined as follows:

$$d^r_{i,j} = \begin{cases} 1 & if d^2_{i,j} >= 1 \\ 0 & d^2_{i,j} = 0 \end{cases} \tag{5.6}$$

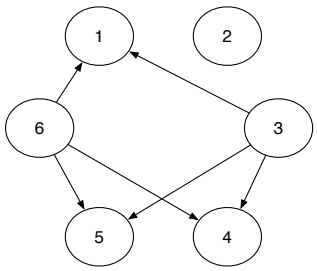

Figure 5.3: Minimum dominance in graph form

	1	2	3	4	5	6
1	0	0	0	0	0	0
2	1	0	0	1	1	0
3	0	1	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	1	0	0	0	0

Table 5.4: Minimum dominance in matrix form

5.2.4 Minimum Cardinality Dominance Matrix

A minimum cardinality dominance matrix, defined as D_m, can be readily defined as:

$$D_m = D_f - D_r$$

Example: By eliminating redundant relationships in the graph, the minimum dominance graph and matrix forms for the example are presented in Figure 5.3.

5.3 Model Comparison and Results

Computation experiments show that the use of minimum cardinality dominance has achieved a dramatic reduction in terms of model size. To see this, Table 5.5 presents the numbers of variables and constraints corresponding to each model. Here, though the number of variables has not changed, the number of constraints has been dramatically reduced, specifically the dominance constraint (5.3), which comprises the largest part of the model. The

original model has a total of 13,833,800 dominance constraints due to the use of a direct or full dominance matrix, the reduced model; however, has only 220,877 constraints.

Model Components		Original Model	Reduced Model
Variables	Allocation (binary) x_i	57,860	57,860
Constraint	One Award per ID (5.2)	5,260	5,260
	Dominance (5.3)	13,833,800	191,497
	Total Budget (5.4)	1	1
	Total Number of Constraints	13,839,061	196,758

Table 5.5: Size of the optimization models

The solution of the original model with a state-of-the-art commercial solver is not possible due to memory limitations; the reduction model, however, was solved in a few minutes on a standard laptop computer (i7-4850HQ 2.3 GHz with 8G of RAM).

5.3.1 Results Under Different SSI

Table 5.6, Table 5.7, and Table 5.8 show the optimization results for SSI set at 10000, 12000 and 14000 respectively. Here "Integer Solution" represents the integer solution value, "Linear Programming" the linear programming solution values, "Optimality Gap" the optimality gap, "# Nodes" the number of nodes in the branch and bound tree, and "Time" the time required in seconds to solve the problem.

SSI	10,000				
Budget	Integer Solution	Linear Programming	Optimality Gap	# Nodes	Time
0.4M	61,516,543	61,527,360	0.02%	0	27
0.6M	61,552,340	61,559,211	0.01%	181	56
0.8M	61,584,196	61,594,905	0.02%	45	78
1.0M	61,603,302	61,616,648	0.02%	0	27
1.2M	61,624,947	61,625,981	0.00%	0	29
1.4M	61,606,891	61,618,654	0.02%	0	108
1.6M	61,593,948	61,602,401	0.01%	402	431
1.8M	61,550,031	61,550,031	0.00%	2270	1080
2.0M	61,552,621	61,557,068	0.01%	0	133
2.2M	61,521,260	61,530,258	0.01%	0	130
2.4M	61,472,929	61,495,934	0.04%	0	145
2.6M	61,428,352	61,455,128	0.04%	80	410
2.8M	61,414,620	61,416,969	0.00%	0	185
3.0M	61,350,350	61,361,072	0.02%	32	209
3.2M	61,260,758	61,272,605	0.02%	1444	640
3.4M	61,185,830	61,197,559	0.02%	598	373

Table 5.6: Computational statistics of optimization model under SSI=10,000

SSI	12,000				
Budget	Integer Solution	Linear Programming	Optimality Gap	# Nodes	Time
1.4M	63,640,276	63,652,166	0.02%	230	341
1.6M	63,656,200	63,667,992	0.02%	31	174
1.8M	63,647,272	63,647,272	0.00%	269	280
2.0M	63,683,212	63,688,055	0.01%	0	48
2.2M	63,685,894	63,692,529	0.01%	0	58
2.4M	63,676,591	63,688,012	0.02%	49	159
2.6M	63,662,739	63,674,086	0.02%	78	250
2.8M	63,668,398	63,669,711	0.00%	0	51
3.0M	63,620,835	63,640,956	0.03%	0	61
3.2M	63,565,618	63,568,716	0.00%	1190	806
3.4M	63,523,681	63,526,856	0.00%	529	323

Table 5.7: Computational statistics of optimization model under SSI=12,000

SSI	14,000				
Budget	Integer Solution	Linear Programming	Optimality Gap	# Nodes	Time
1.4M	65,673,717	65,686,331	0.02%	256	233
1.6M	65,729,171	65,736,548	0.01%	300	202
1.8M	65,743,219	65,745,436	0.00%	1191	470
2.0M	65,815,085	65,819,285	0.01%	0	107
2.2M	65,844,323	65,855,133	0.02%	0	114
2.4M	65,871,923	65,877,439	0.01%	2	224
2.6M	65,885,258	65,902,975	0.03%	0	124
2.8M	65,919,822	65,923,025	0.00%	0	49
3.0M	65,900,599	65,919,720	0.03%	0	58
3.2M	65,871,018	65,874,012	0.00%	808	451
3.4M	65,857,324	65,887,572	0.05%	72	245

Table 5.8: Computational statistics of optimization model under SSI=14,000

The plots of optimization results are shown in Figure 5.4, Figure 5.5 and, Figure 5.6. The horizontal axis represents the budget and the vertical axis the revenue.

As can be seen in Figure 5.4, for SSI = 10,000, a 1.2 million budget yields the maximum revenue return; for SSI = 12,000, a 2.2 million budget yields the maximum revenue;

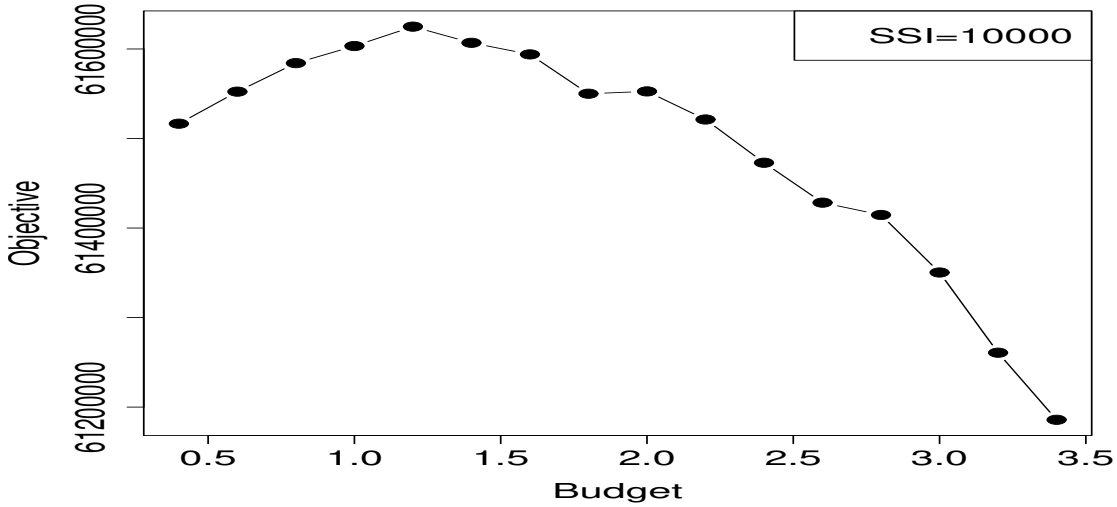

Figure 5.4: Optimization results for SSI = 10,000.

for SSI = 14,000, a 2.8 million budget yields the maximum revenue. In all these figures, revenue increases at the beginning as more financial aid is being allocated and more students are likely to enroll. However, after the point where maximum revenue is reached, additional financial aid negatively affects net tuition and reduces revenue and thus is not desired.

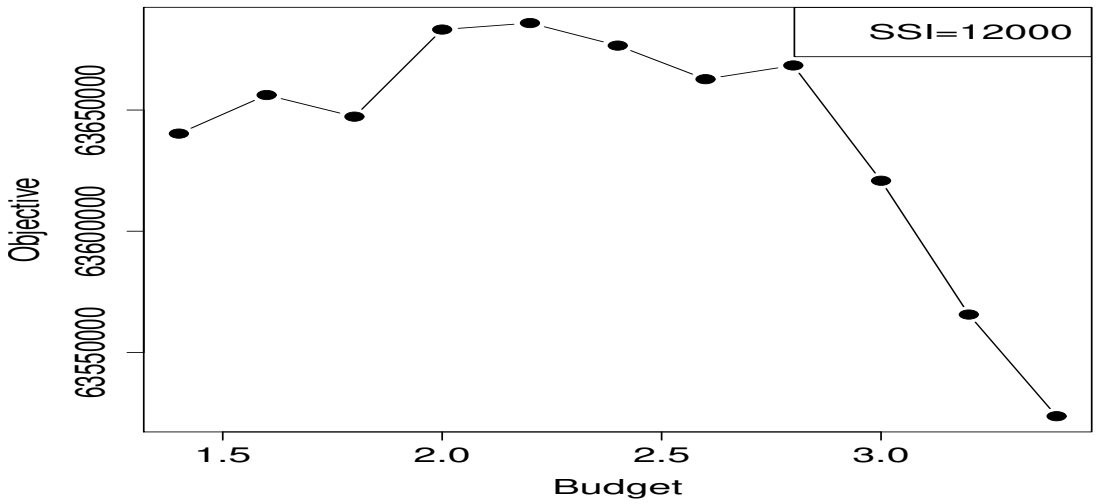

Figure 5.5: Optimization results for SSI = 12,000.

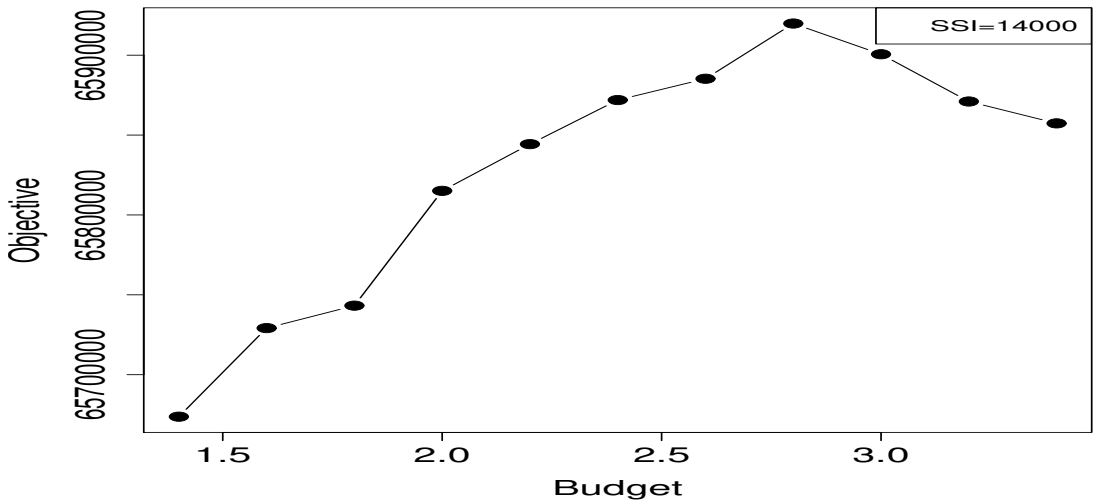

Figure 5.6: Optimization results for SSI = 14,000.

Derivation of Scholarship Award Policies & Implementation

The optimization result based on the above model could be sent to a general linear regression or decision tree for analysis to derive scholarship award policies. These policies represent a simplified solution to the optimal allocation problem; as such, in the development of these policies, it would be desirable that they are simple to understand and do not lose the optimality of allocation.

6.1 Derivation of Scholarship Award Policies

Scholarship award policies can be derived using the above model, which is based on general linear regression or decision tree. Here the predictor or dependent variable is the amount of the award, and the independent variables are GPA, ACT scores, etc. Although variables such as gender could affect the enrollment and graduation probabilities, aid allocation based on these variables is controversial. As a result, in the derivation of financial

aid policies, variables such as gender, family income and ethnicity are not considered in the design of a merit-based scholarship.

6.1.1 Scholarship Award Policy Based on Decision Tree

The decision tree analysis is used in the derivation of financial aid policies. In the past two decades, decision trees, as a decision support tool, have been commonly used in various business domains, such as direct mailing, on-line sales, customer retention and supplier selection, to name a few (Han et al., 2011).

Decision tree analysis is a tree-structured model. There are three types of nodes in a decision tree: a) a root node that has no incoming edges; b) an internal node with one incoming edge and one or more outgoing edges; c) leaf node, which corresponds with a classification rule. Please see (Maimon and Rokach, 2005) for more details.

The financial aid policy initially obtained from the decision tree analysis is presented in Figure 6.1. For example, for a student with GPA = 4.0, ACT = 30, it passes to the right at the root node (ACT < 26.5, no), then to the right at node (GPA <3.85, no), and then to the right at node (ACT <28.5, no), which lands him at a scholarship of $5,339; for for a student with GPA = 3.5, ACT = 28, it passes to the right at the root node (ACT < 26.5, No), then to the left at node (GPA <3.85, yes), and then to the left at node (ACT <29.5, no), then to the right at node (GPA<3,35, no), which lands him at a scholarship of $2,306.

The decision tree analysis, though graphical, still seems a little complicated to reveal the intrinsic patterns of the award.

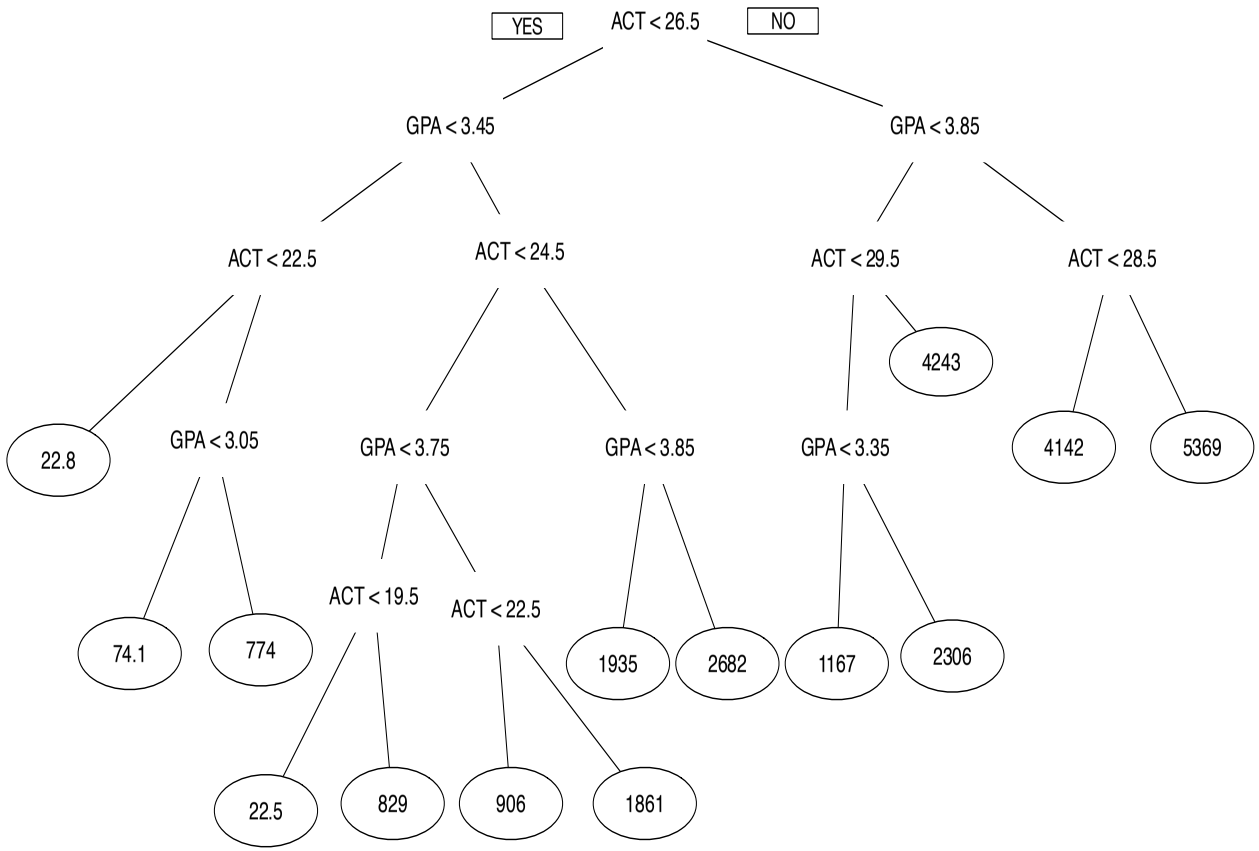

Figure 6.1: Financial aid policy based on decision tree

6.1.2 Scholarship Award Policy on Stepwise Regression

The more intuitive answer to the scholarship allocation from the optimization results is based on two observations; a) the composite score and 2) the piece-wise relations.

1. Composite Score

Table 6.1 presents the average financial aid with respect to the students' GPA and ACT scores. Here the row names represent the GPA scores, and the column names represent the ACT scores. The average financial aid awarded (based on the optimal results with a scholarship budget of $2.4 million) is shown in the corresponding grid.

These results suggest a strong relationship between the awards and both GPA and ACT scores. As a result, a *composite score*, calculated as $10 \times GPA + ACT$ was proposed by the enrollment team to capture the applicants' academic merits and used as the basis for the award.

2. Piecewise Relations

Figure 6.2 shows the piece-wise linear relationship between the average scholarship award and GPA and ACT. Here, the horizontal axis represents GPA and ACT scores and the vertical axis represents the average scholarship awarded for the corresponding score.

Figure 6.2a, for example, clearly shows that no scholarship should be awarded when GPA is below 2.9 or 3.0. In a similar way, no award is allocated for ACT lower than 20 or 21 and for composite score below 57 or 58. A linear relationship seems to exist between the scholarship and the composite score when the composite score is between 55 and 70, and the scholarship seems to remain the same when the score is above 70.

GPA/ACT	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	Total
1																			0
1.1																			0
1.2		0																	0
1.3	0																		0
1.4		0																	0
1.5		0							0										0
2.5	0	0	0	0	0	0	0	0	0	0	0	0		0					0
2.6	0	0	0	0	0	0	0	0	0	0		1250	1500			1500			25.8
2.7	0	0	0	0	0	0	0	0	0	0	1500		2000						24.2
2.8	0	0	0	0	0	0	0	0	0	300	1500	2000				2500			39.4
2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	818.2	1875.0	2000.0								63.4
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1166.7	1500.0	2000.0	2000.0	2200.0		3200.0			5200.0		216.3
3.1	0.0	0.0	0.0	0.0	0.0	0.0	523.8	1500.0	1727.3	2000.0	2000.0	2500.0	2500.0					7300.0	276.7
3.2	0.0	0.0	0.0	0.0	0.0	781.3	1000.0	1750.0	2000.0	2250.0	2500.0	2500.0	2780.0	3950.0	5200.0				573.7
3.3	0.0	0.0	0.0	0.0	647.1	1000.0	1613.6	2000.0	2315.8	2500.0	2920.0	3200.0	3866.7	4200.0		5200.0	5200.0		962.8
3.4	0.0	0.0	0.0	694.4	1000.0	1550.0	2000.0	2000.0	2500.0	2990.0	3200.0	3200.0	4200.0	4200.0					1185.1
3.5	0.0	0.0	695.7	1000.0	1724.1	2000.0	2000.0	2216.7	2500.0	3200.0	3200.0	3200.0	4200.0	4200.0		5200.0			1696.7
3.6	0.0	545.5	1000.0	1000.0	2000.0	2000.0	2357.1	2500.0	2864.0	3200.0	3200.0	3800.0	4200.0	4200.0	5200.0	5200.0	5200.0		2148.1
3.7	0.0	1000.0	1000.0	1500.0	2000.0	2285.7	2500.0	2945.5	3200.0	3200.0	3700.0	4200.0	4200.0	5200.0	5200.0		5200.0		2532.8
3.8	0.0	1000.0	1000.0	2000.0	2000.0	2500.0	3006.9	3200.0	3200.0	4014.8	4200.0	4200.0	4900.0	5200.0	5200.0	5200.0			3007.5
3.9	666.7	1000.0	1000.0	2000.0	2000.0	2500.0	3200.0	3200.0	3680.0	4200.0	4200.0	4644.4	5200.0	5200.0	5200.0	5200.0	5200.0		3360.4
4		1000.0	1000.0	2000.0	2000.0	2500.0	3200.0	3200.0	4200.0	4808.7	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	6542.9	7300.0	4186.1
4.1		1000.0	1000.0	2000.0	2000.0	2500.0	3200.0	3200.0	4200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	7300.0		3982.1
4.2	1000.0		1000.0	2000.0	2000.0	2500.0	3200.0	3200.0	4200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	7300.0		4175.0
4.3					2000.0	2500.0	3200.0	3200.0	4200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	7300.0		4607.9
4.4							3200.0	4200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	5200.0	7300.0		5125.0
4.5			1000.0		2000.0	2925.0		5200.0	5200.0		5200.0	5200.0		5200.0	5200.0	6233.3		7300.0	4605.3
4.6						4200.0			5200.0	5200.0	5200.0		5200.0		5900.0	7300.0	7300.0		5730.0
4.7				5200.0					5200.0	5200.0	6200.0			8400.0	8400.0	8400.0			6377.8
4.8										5200.0	6200.0		6200.0						5866.7
Grand Total	7.2	92.7	166.2	428.7	787.1	1206	1664.8	2112	2642.1	3407.5	3835.6	3981.3	4561.4	4784.9	5323.2	5258.8	6247.6	7300	1134.7

Table 6.1: Optimization mean scholarship vs GPA and ACT

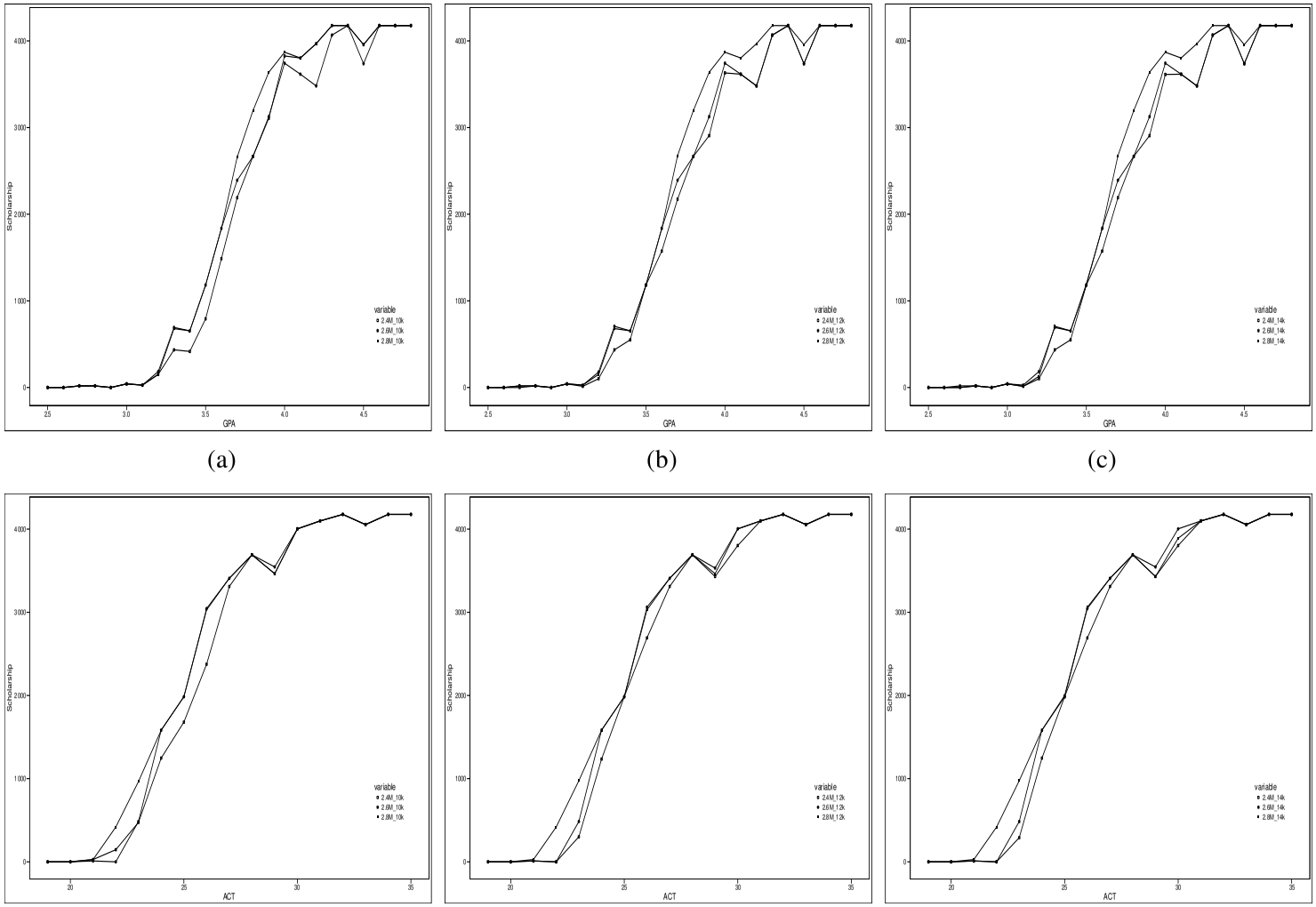

Figure 6.2: (a) (b) (c) Scholarship vs ACT for various budgets and SSI. (d) (e) (f) Scholarship vs GPA for various budgets and SSI.

Stepwise Regression Based on Composite Score

Using "CS" as the variable representing the composite score, a piecewise regression is used to capture these observations and the resulting regression equations are shown in Table 6.2. After the communication with the school enrollment administrations, a simpler discretized version of the scholarship policy is shown in Table 6.3.

Composite Score	# of Applicants	Scholarship Amount
0-53.9	2,897	0
54-68.9	2,103	$309 \times CS - 16,380$
69-76.9	241	$101 \times CS - 2,024$
77-80	19	$711 \times CS - 48,910$

Table 6.2: Piecewise scholarship allocation

Composite Score	Scholarship Amount
0-54.9	0
55-59.9	1,500
60-65.9	2,500
66-69.9	3,500
70-74.9	4,500
75+	6,000

Table 6.3: Discretized version of scholarship allocation

6.1.3 Insights on Change Of Budget

Figures 6.3, 6.4, and 6.5 show the relationship between average scholarship award and GPA and ACT for different levels of budget and different levels of SSI. Here, the horizontal axis represents composite scores, the vertical axis represents average corresponding scholarship awarded, different regression lines represent different total budgets, and different figures represent different levels of SSI.

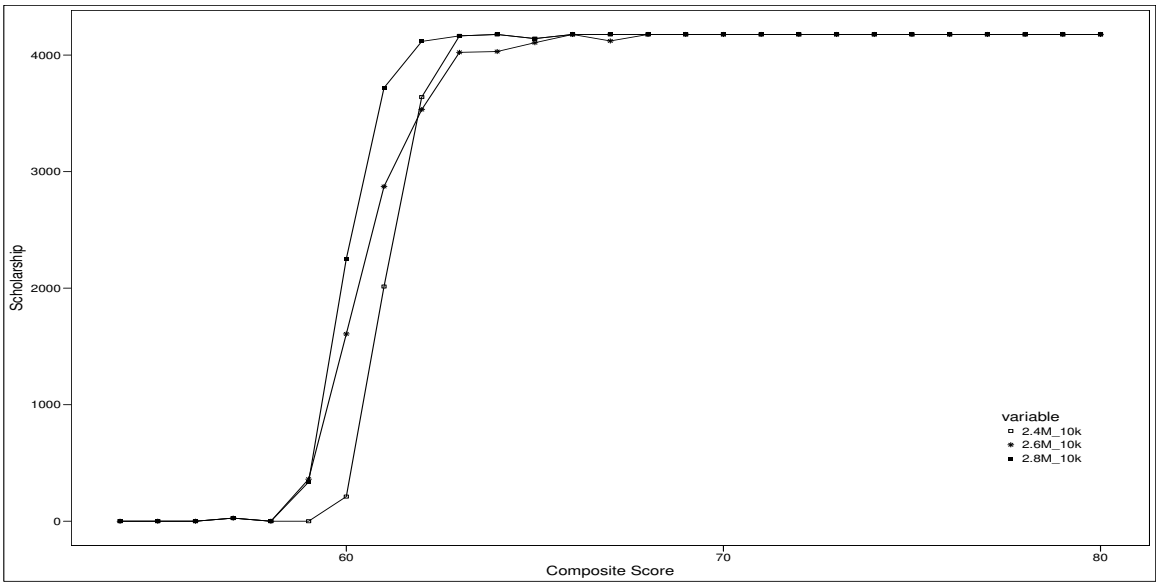

Figure 6.3: Scholarship vs Composite score for SSI=10,000

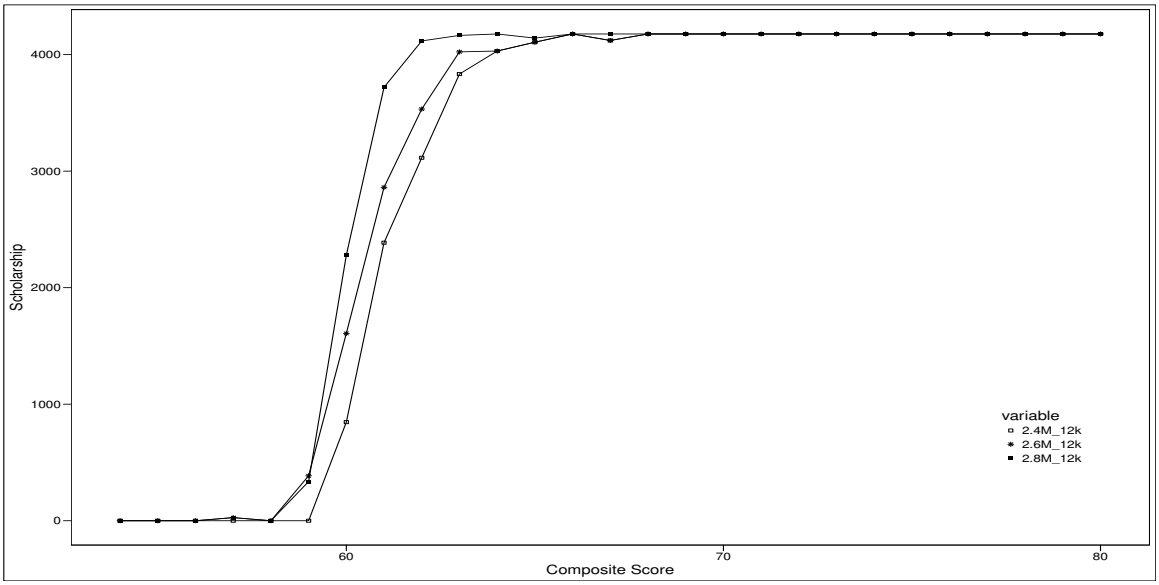

Figure 6.4: Scholarship vs Composite score for SSI=12,000

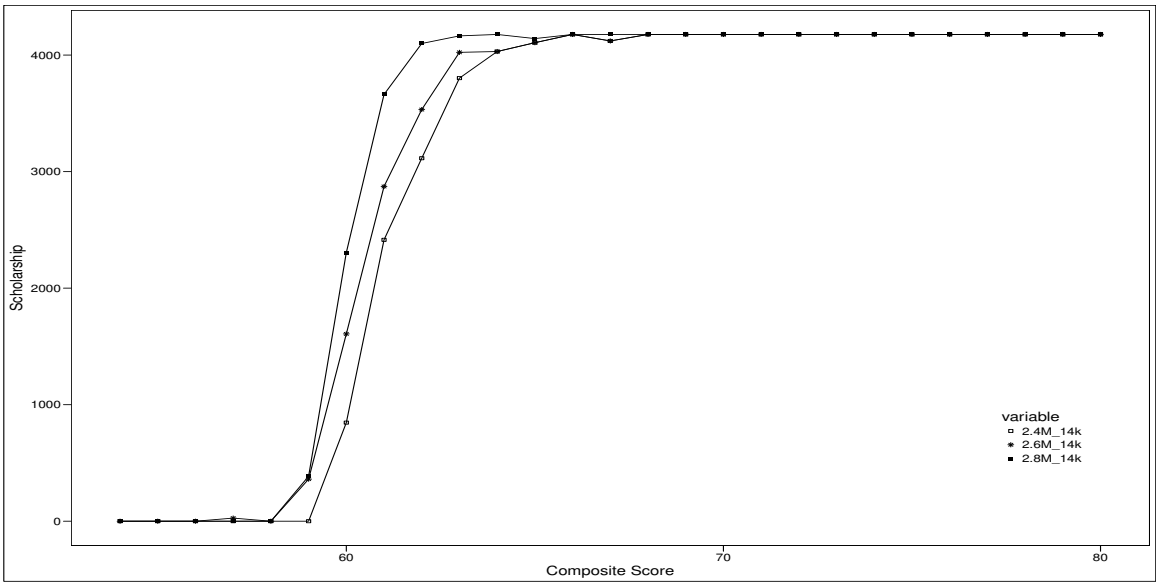

Figure 6.5: Scholarship vs Composite score for SSI=14,000

6.2 Implementation and Results

The scholarship allocation based on the composite score and the policy presented in Table 6.2 has been used as the foundation for the scholarship for the university in the 2013 to 2014 academic year. The university has taken a proactive approach. The enrollment and admission office has purchased data on student performance and scholarship awards before they even apply (these awards hinge upon the verification of their official performance).

Table 6.4 presents the enrollment statistics for the university in the 2012 - 2013 academic year and those of the 2013 - 2014 academic year.

	2013	2014	# Increase	% Increase
Application	6,101	6,068	-43	-0.7%
Admitted	4,541	4,773	232	5.1%
Non-Scholarship	2,166	2,157	-9	-0.4%
Scholarship Award	2,375	2,616	241	10.1%
Matriculated	2,001	2,222	221	11.0%

Table 6.4: Enrollment statistics for the 2012-2013 and 2013-2014 academic years

In 2012-2013, there are a total of 6,101 applicants; 4,541 were admitted, 2,375 were awarded scholarships, and 2,166 were not awarded scholarships. 52% of the students were awarded scholarships, and a total of 2,001 students matriculated.

In the 2013-2014 academic year, there were a total of 6,068 applicants; of them, 4,773 were admitted, 2,616 were awarded scholarships, and 2,157 were not awarded scholar-

ships. 56% of the students are awarded scholarships, and a total of 2,222 students matriculated.

Notice that the number of applicants does not change dramatically, actually showing a reduction of -43 (0.7% decrease), but the actual enrollment increased by 221 or 11.0% over the previous years. It is believed that the use of the optimal policy could generate millions of dollars in revenue for the university in the next few years.

Table 6.5 shows the optimization model results under different level.The objective reaches highest at 2.2 million budget when SSI = 12,000. When spending 0 and 2.2 million on budget, the average graduation probability is 43% and 45% respectively. The raw revenue of spending 0 (6.1) and 2.2 (6.2) million are calculated as:

$$2167 * 8354 * 0.43 * 4 + 2167 * 0.43 * 12000 = 42,319,083 \tag{6.1}$$

where 2167 is the number of students expected to enroll, 8,354 is the tuition, and 12,000 is the SSI.

$$2356 * (8354 - 933) * 0.45 * 4 + 2356 * 0.45 * 12000 = 44,193,377 \tag{6.2}$$

where 933 is the average scholarship of the each student gets under 2.2 million budget.

As a result, the profit of spending extra 2.2 million is 44,193,377 - 42,319,083 = 1,874,294.

budget	objective	num_student
0	63,222,695	2,167
0.2	63,303,024	2,185
0.4	63,385,102	2,203
0.6	63,449,801	2,221
0.8	63,518,563	2,239
1	63,578,064	2,256
1.2	63,619,230	2,273
1.4	63,640,276	2,290
1.6	63,656,200	2,307
1.8	63,647,273	2,322
2	63,683,212	2,340
2.2	63,685,894	2,356
2.4	63,676,591	2,372
2.6	63,662,739	2,387
2.8	63,668,398	2,404
3	63,620,835	2,418
3.2	63,565,619	2,433
3.4	63,523,681	2,448

Table 6.5: Optimization results when SSI=12,000

Conclusion

This research studies the optimal financial aid allocation problem that puzzles many higher education institutions. The problem is complex yet of financial importance. Various techniques have been investigated and a three-phase framework was proposed in the paper as the solution of the optimal scholarship allocation problem to derive simple yet effective financial aid policies.

In the first phase, a series of predictive models have been investigated to estimate two types of responses from students with financial aid awards. The first response is enrollment and graduation decisions from students with various socioeconomic characteristics; the second response is the number of years of study once a student enrolls in the institution.

In the first case, because of the binary nature of the responses, "enroll" or "not enroll", "graduate" or "not graduate", logistic regression based models have been adopted to predict the probability of enrollment of an applicant and the probability of graduation given that he/she enrolls. In the second case, a regression analysis is to be used to predict the number of years of study once the student enrolls in the institution.

In the second phase, an integer linear program is designed to allocate financial aid to applicants with an objective to maximize the revenue, which is composed of tuition minus

scholarship allocated over the years, plus the state share of instruction (SSI) once the student graduates. The constraints to be observed include the total budget limitations as well as other considerations such as fairness. For a merit-based scholarship, the fairness constraint stipulates that a student with better academic performance should be assigned to an equal or higher level of scholarship than that of a students with a lower academic performance. The inclusion of the fairness constraints has dramatically increased the size of the model and a model reduction technique, referred to as minimum cardinality dominance, had to be developed to solve the model effectively.

A computational experiment shows that the use of minimum cardinality dominance has achieved a dramatic reduction regarding model size. In a test case, pairwise comparison of 6,000 students was reduced from more than 13.5 million constraints to only 191,000 constraints, enabling effective solution of the models. In this particular case, the original model is computationally unsolvable, actually running out of memory because of the large model size; the reduced model nevertheless can be solved in minutes.

In the third phase, regression analysis is developed to translate the optimization results, in the form of the amount of scholarship awarded for each student, into managerial insights and to derive a policy for implementation. The analysis suggested that the use of a composite score, derived based on the student's GPA and ACT scores, could be used as the basis in the award of scholarships to form simple yet effective scholarship policies.

The result of the study has been successfully implemented in the exemplary state university and has resulted in millions of financial benefits. The research would be applicable to many other institutions and offers methodology, tools and insights into the solution of financial aid problems.

Future Studies. The results from the optimization specify the scholarship awards to each applicant under a specific population and budget. The actual size and composition of the application pool could be affected by the unemployment rate and is random in nature. Nevertheless, at this stage, this study is on optimization under a specific pool. Stochastic optimization techniques such as sampling to find the optimal allocation under a random pool will be of great interest.

Bibliography

K. G. Abraham and M. A. Clark. Financial aid and students' college decisions: Evidence from the district of columbia tuition assistance grant program. *The Journal of Human Resources*, 41(3):578–610, 2006.

Anthony J. Adam and Gerald H. Gaither. Retention in higher education: A selective resource guide. *New Directions for Institutional Research*, 2005(125):107–122, 2005.

S.S. Aksenova, D. Zhang, and M. Lu. Enrollment prediction through data mining. In *2006 IEEE International Conference on Information Reuse and Integration*, pages 510–515, Sept 2006.

B. L. Bailey. Let the data talk: Developing models to explain ipeds graduation rates. *New Directions for Institutional Research, Special Issue: Data Mining in Action: Case Studies of Enrollment*, 2006(131):101–115, 2006.

A. Belloni, M. J. Lovett, W. Boulding, and R. Staelin. Optimal admission and scholarship decisions: Choosing customized marketing offers to attract a desirable mix of customers. *Marketing Science*, 31(4):621–636, 2012.

D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons, 2005.

M.D Borah, R. Jindal, D. Gupta, and G. C. Deka. Application of knowledge based decision technique to predict student enrollment decision. In *2011 Recent Trends in Information Systems (ReTIS) on International Conference*, pages 180–184, Dec 2011.

A. Braunstein, M. Mcgrath, and D. Pescatrice. Measuring the impact of income and financial aid offers on college enrollment decisions. *Research in Higher Education*, 40(3): 247–259, 1999.

T.H. Bruggink and V. Gambhir. Statistical models for college admission and enrollment: A case study for a selective liberal arts college. *Research in Higher Education*, 37(2): 221–240, 1996.

R. E. Carter and D. J. Curry. Using student-choice behaviour to estimate tuition elasticity in higher education. *Journal of Marketing Management*, 27(11-12):1186–1207, 2011.

B. L. Castleman and B. T. Long. Looking beyond enrollment the causal effect of need-based grants on college access, persistence, and graduation. *The Journal of Labor Economics*, 34(4):1023–1073, 2016.

K. Y. Chan and W. Y. Loh. Lotus: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, 13(4): 826–852, 2004.

T. Chang. Data mining: A magic technology for college recruitment. *Paper of Overseas Chinese Association for Institutional Research (www.ocair.org)*, 2008.

R.G. Chapman and R. Jackson. *College choices of academically able students: the influence of no-need financial aid and other factors*. Research monograph. College Entrance Examination Board, 1987.

W. Chen, J. Song, L. Shi, L. Pi, and P. Sun. Data mining-based dispatching system for solving the local pickup and delivery problem. *Annals of Operations Research*, 203(1): 351–370, 2013.

S. Cohodes and J. S. Goodman. Merit aid, college quality, and college completion: Massachusetts' adams scholarship as an in-kind subsidy. *American Economic Journal: Applied Economics*, 6(4):251–283, 2014.

J. T. Crouse. Estimating the average tuition elasticity of enrollment for two-year public colleges. *American Journal of Economics*, 5(3):303–314, 2015.

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

Carolyn E Cutrona, Valerie Cole, Nicholas Colangelo, Susan G Assouline, and Daniel W Russell. Perceived parental social support and academic achievement: an attachment theory perspective. *Journal of personality and social psychology*, 66(2):369, 1994.

M Scott DeBerard, Glen Spielmans, and Deana Julka. Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College student journal*, 38(1):66–80, 2004.

Gerben Dekker, Mykola Pechenizkiy, and Jan Vleeshouwers. Predicting students drop out: A case study. In *Educational Data Mining 2009*, 2009.

Stephen L. DesJardins, Dennis A. Ahlburg, and Brian P. McCall. An integrated model of application, admission, enrollment, and financial aid. *The Journal of Higher Education*, 77(3):381–429, 2006.

B. Donald, L. Lichtenstein, G. Palumbo, and M. P. Zaporowski. Optimization techniques for college financial aid managers. *Journal of Student Financial Aid*, 40(3), 2010.

S. M. Dynarski. Hope for whom? financial aid for the middle class and its impact on college attendance. *National Tax Journal*, 53(3):629–662, 2000.

S. M. Dynarski. The consequences of lowering the cost of college: The behavioral and distributional implications of aid for college. *American Economic Review*, 92(2):279–285, 2002.

S. M. Dynarski. Does aid matter? measuring the effect of student aid on college attendance and completion. *The American Economic Review*, 93(1):279–288, 2003.

S. M. Dynarski and J. Scott-Clayton. Financial aid policy: Lessons from research. *The Future of Children*, 23(1):67–91, 2013.

R. G. Ehrenber. Econometric studies of higher education. *Higher Education*, 1-2:19–37, 2004.

R. G. Ehrenberg and D. R. Sherman. Optimal financial aid policies for a selective university. *The Journal of Human Resources*, 19(2):202–230, 1984.

Jerome Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367 – 378, 2002.

M. B. Fuller. A history of financial aid to students. *Journal of Student Financial Aid*, 44 (1), 2014.

M. Habshah, S. K. Sarkar, and S. Rana. Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, 13(3):253–267, 2010.

J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

F. E. Harrell. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Science & Business Media, 2013.

D. E. Heller. Student price response in higher education: An update to leslie and brinkman. *The Journal of Higher Education*, 68(6):624–659, 1997.

D. E. Heller. The effects of tuition and state financial aid on public college enrollment. *The Review of Higher Education*, 23(1):65–89, 1999.

S. Herzog. Estimating student retention and degree completion time: Decision trees and neural networks vis-a-vis regression. *New Directions for Institutional Research, Special Issue: Data Mining in Action: Case Studies of Enrollment*, 131, 2006.

D. E. Hinkle, W. Wiersma, and S. G. Jurs. *Applied statistics for the behavioral sciences*. JSTOR, 2003.

Louise. Horstmanshof and Craig. Zimitat. Future time orientation predicts academic engagement among first-year university students. *British Journal of Educational Psychology*, 77(3):703–718, 2007.

D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 2013.

D. Hossler and J. P. Bean. *The Strategic Management of College Enrollments*. Jossey-Bass, 1990.

D. Hossler, J. Braxton, and G. Coopersmith. Understanding student college choice. *Higher education: Handbook of theory and research*, 5:231–288, 1989.

D. Hossler, J. Schmit, and N. Vesper. *College Choice: Understanding Student Enrollment Behavior. ASHE-ERIC Higher Education Report No. 6.* Johns Hopkins University Press, 1998.

G. Jackson. Financial aid and student enrollment. *The Journal of Higher Education*, 49 (6):548–574, 1978.

Gregory A. Jackson. Did college choice change during the seventies? *Economics of Education Review*, 7(1):15 – 27, 1988.

F. R. Kemerer, J. V. Baldridge, and K. C. Green. *Strategies for effective enrollment management*. American Association of State Colleges and Universities, 1982.

D. Kim. The effect of financial aid on students' college choice: Differences by racial groups. *Research in Higher Education*, 45(1):43–70, 2004.

Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer Science & Business Media, 2008.

Zlatko J. Kovačić. Early prediction of student success: Mining student enrollment data. In *Proceedings of Informing Science & IT Education Conference*, 2010.

Rabby Q. Lavilles and Mary Jane B. Arcilla. Enrollment forecasting for school management system. *International Journal of Modeling and Optimization*, 2:563–566, 2012.

L. L. Leslie and P. T. Brinkman. Student price response in higher education: The student demand studies. *The Journal of Higher Education*, 58(2):181–204, 1987.

L. L. Leslie and P. T. Brinkman. *The Economic Value of Higher Education. American Council on Education/Macmillan Series on Higher Education.* Collier Macmillan Publishers, 1988.

Joe J. Lin, Kenneth J. Reid, and P. K. Imbrie. Work in progress - predicting retention in engineering using an expanded scale of affective characteristics from incoming students. In *Proceedings of the 39th IEEE International Conference on Frontiers in Education Conference*, FIE'09, pages 616–617. IEEE Press, 2009.

J Scott Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*. Stata press, 2006.

J. Maguire. To the organized go the students. *Bridge Magazine*, 39(1):16–20, 1976.

Oded Maimon and Lior Rokach. *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag New York, Inc., 2005.

E. N. Maltz, K. E. Murphy, and M. L. Hand. Decision support for university enrollment management: Implementation and experience. *Decision Support Systems*, 44(1):106–123, 2007.

Dolores W Maney. Predicting university students' use of alcoholic beverages. *Journal of College Student Development*, 31(1):23–32, 1990.

M. K. McLendon, D. A. Tandberg, and N. W. Hillman. Financing college opportunity: Factors influencing state spending on student financial aid and campus appropriations from 1990 to 2010. *The ANNALS of the American Academy of Political and Social Science*, 665(1):143–162, 2014.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017. R package version 1.6-8.

Diane Musgrave-Marquart, Susan P Bromley, and Mahlon B Dalley. Personality, academic attribution, and substance use as predictors of academic achievement in college students. *Journal of Social Behavior and Personality*, 12(2):501, 1997.

National Center for Education Statistics. Mobile digest of education statistics. https://nces.ed.gov/programs/digest/mobile/Finance_Degree-Granting_Institutions_Revenues_for_Public_Institutions.aspx, 2014. [Accessed: 2017-09-21].

Kimberly Noble, Nicole T. Flynn, James D. Lee, and David Hilton. Predicting successful

college experiences: Evidence from a first year retention program. *Journal of College Student Retention: Research, Theory & Practice*, 9(1):39–60, 2007.

M. B. Paulsen. *College Choice: Understanding Student Enrollment Behavior. ASHE-ERIC Higher Education Report No. 6.* Ashe-Eric Higher Education Reports, 1990.

C. Peng, T. So, F. Stage ., and E.P. John. The use and interpretation of logistic regression in higher education journals: 1988–1999. *Research in Higher Education*, 43(3):259–293, 2002.

Vahe Permzadian and Marcus Credé. Do first-year seminars improve college grades and retention? a quantitative review of their overall effectiveness and an examination of moderators of effectiveness. *Review of Educational Research*, 86(1):277–316, 2016.

MN Quadri and NV Kalyankar. Drop out feature of student data for academic performance using decision tree techniques. *Global Journal of Computer Science and Technology*, 10(2), 2010.

Werner J. Reinartz and V. Kumar. The impact of customer relationship characteristics on profitable lifetime duration. *Journal of Marketing*, 67(1):77–99, 2003. ISSN 00222429.

Greg Ridgeway. *gbm: Generalized Boosted Regression Models*, 2017. R package version 2.1.3.

N. Seftor and S. Turner. Back to school: Federal student aid policy and adult college enrollment. *The Journal of Human Resources*, 37(2):336–353, 2002.

D. L. Sjoquist and J. V. Winters. State merit-based financial aid programs and college attainment. *The Journal of Regional Science*, 55(3):364–390, 2015.

P. K. Sugrue. An optimization model for the allocation of university based merit aid. *Journal of Student Financial Aid*, 40(2), 2010.

Patrick T. Terenzini, Wendell G. Lorang, and Ernest T. Pascarella. Predicting freshman persistence and voluntary dropout decisions: A replication. *Research in Higher Education*, 15(2):109–127, 1981.

L. V. Thanh and P. Haddawy. Deriving financial aid optimization models from admissions data. In *Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE '07. 37th Annual*, pages F2A–7–F2A–12, Oct 2007.

Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125, 1975.

Vincent Tinto. Limits of theory and practice in student attrition. *The Journal of Higher Education*, 53(6):687–700, 1982.

Rajkumar Venkatesan and V. Kumar. A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, 68(4):106 – 125, 2004. ISSN 00222429.

Eric-Jan Wagenmakers and Simon Farrell. Aic model selection using akaike weights. *Psychonomic bulletin & review*, 11(1):192–196, 2004.

T. Yamashita, A. J. Bailer, D. A. Noe, T. Yamashita, A. J. Bailer, and D. A. Noe. Identifying at-risk subpopulations of canadians with limited health literacy. *Epidemiology Research International*, 2013.

Chong Ho Yu, Samuel DiGangi, Angel Jannasch-Pennell, and Charles Kaprolet. A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, 8(2):307–325, 2010.

Ying Zhang, Samia Oussena, Tony Clark, and Kim Hyensook. Using data mining to improve student retention in he: a case study. In *ICEIS 2010, 12th International Conference on Enterprise Information Systems*, 2010.