Leveraging large-scale educational data to examine the freshman experience using machine learning and causal inference methods

Lovenoor Singh Aulck

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Jevin West, Chair

Emma Spiro

Joshua Blumenstock

Program Authorized to Offer Degree:
Information Science

University of Washington

Abstract

Leveraging large-scale educational data to examine the freshman experience using machine learning and causal inference methods

Lovenoor Singh Aulck

Chair of the Supervisory Committee:
Assistant Professor Jevin West
The Information School

Institutions of higher education are constantly collecting student-centric data, be it application information, transcript records, or graduation histories. Despite this wealth of data, traditional institutions, where learning is primarily on-campus and in-person, are often limited by infrastructure and personnel in their ability to transform their data into useful information to make data-informed decisions. In this work, I employ supervised machine learning and econometric techniques to utilize data that is routinely collected at traditional institutions of higher education to improve institutional processes and better understand students. First-year, first-time students face numerous challenges as they transition to post-secondary education and, as a case study, I examine two phases of their academic careers: their enrollment decision prior to starting their college education and their first year on campus as they adjust to college life and push towards graduation. In all, this work demonstrates how academic institutions can apply data-centric techniques and contributes to long-standing education theory on post-secondary enrollment, acclimation, and persistence.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my advisor, Jevin West, who was a constant source of optimism and positivity throughout my doctoral experience. I would also like to thank Emma Spiro and Joshua Blumenstock, who, alongside Jevin, established the DataLab at the University of Washington with a central theme of applied data science research with a strong social good focus. The three of them, through interactions both large and small, helped direct my doctoral work towards the direction it ultimately took. I would also like to thank the remaining DataLab faculty and DataLab students for helping foster a positive and productive research environment. In addition, I would like to thank Dan Ratner and Jesse Burk-Rafel, who laid much of the groundwork upon which my work was ultimately built. I would also like to thank those that provided financial support for my work including the National Science Foundation, the Allen Institute for Artificial Intelligence, the University of Washington's iSchool, and various other University of Washington departments and administrative units. Lastly, I would like to thank my friends and family, who are always a constant source of strength and inspiration.

DEDICATION

To GSK, who would always ask when I was going to have "DR." in front of my name.

Chapter 1

INTRODUCTION

1.1 The Irony

When I began my PhD at the University of Washington's (UW's) Information School (iSchool), my intention was to conduct healthcare research using large-scale data. The specifics of this research were rather murky, as they so often are for newly-admitted PhD students, but I was fairly certain that I wanted to investigate problems in healthcare. My background was in bioengineering and I had been working on public/occupational health-related research after graduating with my first Master's degree. Continuing on to a PhD while conducting research in this same space seemed like a natural progression. I had even pitched ideas around healthcare informatics to faculty during my interview for admission to the iSchool Information Science PhD program. After my acceptance to the program, I thought having Jevin West, with his background in theoretical biology, as a faculty mentor would also help with these pursuits. Of course, these pursuits never materialized. Instead, during my first quarter as a PhD student, Jevin suggested I begin with some research by exploring a dataset he had while we thought more about problems we could tackle in healthcare. Suffice it to say, I began digging through the dataset he shared, went down the proverbial rabbit hole, and never turned back.

The dataset Jevin shared comprised nearly the entirety of UW's Registrar data and had been compiled in conjunction with researchers from UW's Bioengineering Department (who were, ironically, my former undergraduate professor, Dan Ratner, and TA, Jesse Burk-Rafel). It included data on student transcripts, demographics, registration histories, course/instructor evaluations, and degrees awarded across the entirety of UW's 150+ year history for which data was digitized. It had originally been compiled in response to Jevin

and Dan's frustrations with the lack of information on students who were in their classes and departments. Who were these students? What were their backgrounds? What were they learning and what did they already know? These questions dug much deeper than the mere list of names on the course roster provided to instructors at the start of an academic term. When I began looking through the data, my own questions naturally began to emerge, colored by my own experiences as a former undergraduate at UW. Which students were most at risk of dropping out? Which course-taking patterns were most effective for students to graduate on time? How did instructor quality impact student success? Of course, not all of these were ultimately answered and I leave the UW still with enough questions to complete a dissertation several times over. Some of the questions that were answered, in whole or in part, are what comprise this dissertation.

As I poured through the data and answered questions as I could, I also became more involved with the administrative circles that managed and used this data. Soon, an oddity began to emerge: within these administrative circles, the data I was working with was predominantly used for record-keeping and not for research or analysis. In the age of "big data", the UW had a wealth of data on its institutional history, the students that have come through its doors, and its academic ecosystem yet this data wasn't being leveraged for any insights. The irony of this becomes more apparent when considering UW's standing as a leader in data science research and innovation, from the tens of millions of dollars invested into its eScience Institute to proclamations of being "Big Data U" with a "university-wide approach" to data science education[1]. Yet, despite its place at the forefront of *developing* data science, the UW was lagging in terms of *applying* what it was teaching to its own data. I came to realize that this was not unique to UW but true of most institutions of higher education. Universities the world over were collecting data on students and instructors as a matter of record-keeping protocol only to have it gather digital dust over time. Of course, the reasons behind this are numerous, from a lack of trained personnel to bureaucratic overhead

[1]See: http://data.washington.edu/ and https://escience.washington.edu/education/. As of April 10, 2019, these phrases were prominently placed on these UW webpages.

to a lack of usable digitized data. Regardless, the fact remained - universities were developing data science but they weren't using it internally like they could.

This is what ultimately pushed me to complete my dissertation within the space of educational data mining, with a particular emphasis on using data that is procedurally collected at traditional[2] institutions of higher education. The aim of this dissertation is to help resolve the irony. It is to demonstrate how data science *at* a university can be used *for* a university. It is to demonstrate how data that is routinely collected at institutions of higher education can be used to improve and learn more about these same institutions.

1.2 EDM

Educational Data Mining (EDM) is a relatively new field that focuses on data-intensive methods and analysis in the realm of education [156, 37, 137]. The growth of the field has been spurred by two (broad) key factors [17]: 1) an increased availability, digitization, and standardization of education-related data and 2) a growth in computational and software capacity to draw insights from this data. The first relates primarily to the increase in education-centric data over the last decade or so, particularly by way of online learning environments, digitized student transcript/registrar records, and more robust mechanisms to digitally capture student activity (card swipes, food service accounts, etc.) [17, 155]. In many respects, this has allowed for the quantification/monitoring of spaces within education that were previously unquantifiable until the last few decades. The second point that has helped spur the growth of EDM relates to how the increased performance of computational machines (as highlighted by Moore's Law [118]) and lower costs for data and hardware storage/maintenance (particularly with respect to the cloud) are continually lowering technology barriers and increasing computational accessibility [30]. It should also be no surprise that these two factors which have influenced advances in EDM - namely, an increased ability to collect data (via a proliferation of data sources and the ubiquity of collection tools) and an

[2] "Traditional" in the sense that the primary mode of instruction is on campus, in person and not online.

increased capacity to store/analyze it - are those that have helped catalyze the growth of "data science" at large, particularly over the last two decades [60, 30, 79].

Often, the very fact that the phrase "data mining" exists in EDM's name causes some consternation. In particular, data mining's association with customer data, advertisers, government intrusion, and ever-watchful eyes is often pejorative [105], justifiably or not, and flies in the face of education's historical goal of remaining a noble pursuit. Adding to this, education as a field is notoriously slow to change to new trends and incorporate new technologies despite having an increasing abundance of data [147, 26]. That is all to say that the growth of EDM, despite its promise [26, 19, 137], faces obstacles with regards to its position within the broader scope of education. These obstacles are not only analogous to any resistance from social science fields towards the adoption of more computation-centric approaches but are also specific to education with respect to slow adaptation overall and a generally risk-averse climate [119]. These obstacles are even more prevalent in post-secondary education, where leveraging data for insights is less common at traditional campuses (i.e. schools where learning is primarily on-campus and in-person) and more expansive in online and computerized environments, which are much more amenable to the collection and analysis of digitized student traces [125]. To this end, traditional universities remain "data-rich" but are "information-poor" in that they have the raw data needed to extract intelligible insights but are unable to do so due to infrastructure limitations, untrained personnel, and a general aversion to risk taking [159].

The focus of my dissertation is to demonstrate how data that is routinely collected at traditional institutions of higher education can be leveraged for insights using supervised machine learning and causal inference methods. To do so, I use first-time, first-year (freshmen) students as a case study. I hope this dissertation will accomplish the following:

- Firstly, I hope this work will be used to directly shape policy at UW. The work I present in this dissertation was completed at UW using data on students at UW. The research can be used to better institutional processes and directly inform decisions at

the University.

- Secondly, I hope the chapters in this dissertation stand as case studies for how other institutions of higher education can apply data-centric tools/approaches to their historical records and gain insights to improve how they recruit, monitor, and support students. Universities are hubs for data-driven research and instruction; yet their ability to apply these data-centric tools to their own administrative data is underdeveloped. I want my work to help change that.

- Thirdly, I hope this dissertation contributes to existing education theory with my findings. This includes contributing to literature on: the impact of financial aid on student enrollment, influences on student persistence towards graduation, and the impact of co-enrollment seminars to help students acclimate to the college environment. Rather than relying on approaches previously used in education research, I either use data-centric methodologies that are underutilized in education research, examine students at a scale that is uncommon in education research, or both.

1.3 Description of Chapters

This dissertation is divided into two Sections, each of which contain two substantive Chapters. Section I focuses on freshmen students prior to their enrollment at the University. It examines the effects of scholarships on student enrollment and optimizing the allocation of merit-based awards. Section II focuses on freshmen after their enrollment at the University. It looks at predicting whether they will eventually graduate and the effect of freshmen seminar classes on their performance. Of the two Chapters in each Section, one focuses on predictions using supervised machine learning (Chapters II and IV) and the other focuses on making causal inferences using econometric approaches like propensity score matching and regression discontinuity designs (Chapters III and V).

Each Chapter consists of a standalone research project that has either been published, is submitted for publication, or is being readied for publication. Each Chapter is relatively

identical in form to how it was published or will be published - each has its own abstract, introduction, methods, results, and conclusions subsections. For cohesiveness, I've also included a short preamble to each of the Chapters that gives a brief overview of the work in the context of other dissertation Chapters. At the end of each chapter, I've also included a list of relevant published/presented work that I've completed related to the Chapter. In all, the work presented in these Chapters draws on a breadth of approaches from the realms of supervised machine learning (e.g. binary classification, algorithmic calibration, numerical optimization) as well as econometrics (e.g. regression, regression discontinuity, propensity score matching).

The Chapters and Sections of this dissertation are organized as follow:

- *Section I: Examining freshmen enrollment*
  - *Chapter II: Predicting student enrollment and optimizing scholarship disbursement*

    This chapter centers on using machine learning and genetic algorithms to optimize a scholarship fund and examining how enrollment management can be improved using data. I first built a machine learning model to predict freshmen enrollment using data from student applications. I then leveraged this predictive model in conjunction with a genetic algorithm that altered scholarship award amounts to determine how to maximize predicted enrollment from the scholarships. After optimization, the scholarships were disbursed to an incoming class of students and the real-world enrollment numbers from this class are compared to previous classes. I found that the optimized approach increased enrollment yield for the university. The approach has since been incorporated into the University's scholarship disbursement pipeline.
  - *Chapter III: Understanding the effect of scholarships on the student enrollment decision*

    This chapter centers on investigating the optimized allocations from *Chapter I* to

understand the per-dollar impact on students' enrollment decisions. I first used data from previous scholarship disbursements to examine the per-dollar impact of merit-based scholarships on student enrollment and the degree to which this could be effectively evaluated. I then compared those who did and did not receive scholarships from *Chapter I* using a regression discontinuity design. I go through multiple approaches in estimating the effect of scholarships on student enrollment. I find some evidence that smaller scholarships awarded to students who have lower academic performance tend to have a greater impact than larger scholarships awarded to students with higher academic performance.

- *Section II: Examining freshmen persistence*
  - *Chapter IV: Predicting first-year undergraduate attrition*

    This chapter centers on using machine learning approaches to understand when and why students leave the university after their first-year. I built machine learning classifiers to predict freshman attrition using only a single year's worth of data on each student. I did this across the entirety of the UW's student body and not just a subset thereof. I examined attrition across two commonly used definitions from the literature: whether a student returns for their second year and whether a student eventually graduates with a degree. I also explored how well subsets of different institutional data fare in the prediction tasks in isolation. I found that using students' first year of transcript records, attrition can be accurately predicted. Additionally, I also found that transcript-based features fared far better than demographic and pre-entry features in predicting attrition.

  - *Chapter V: Understanding the effect of freshmen seminars on undergraduate performance*

    This chapter centers on examining a freshmen seminar program that has been in existence for over 30 years called First-Year Interest Groups (FIGs). I examined its effect on student grades and persistence, including which aspects of FIGs students

find most beneficial. I used matched comparison groups based on propensity score matching to compare FIG and non-FIG students in terms of graduation rates and grades. I also performed this same comparison for particular subsets of students, namely Hispanic and under-represented students. In addition to using registrar data to develop these matched comparison groups, I also examined which aspects of FIGs were most beneficial to students using student exit survey data. To do this, I used topic modeling across the surveys to develop a preliminary codebook of student responses before hand-coding individual student responses. I found that completing a FIG has a positive effect on student grades and persistence. I also found that students found social aspects of the FIGs to be the most beneficial to their educational experience.

The above four Chapters comprise the substantive research Chapters of this dissertation. In addition, there is also this Introduction Chapter (Chapter I), a Conclusions Chapter (Chapter VI), as well as references at the end of this manuscript.

Chapter 2

PREDICTING STUDENT ENROLLMENT AND OPTIMIZING SCHOLARSHIP DISBURSEMENT

2.1 PREAMBLE

In this Chapter, I examine students as they are deciding on admissions offers. I focus on the degree to which we can predict whether a student will enroll and on finding optimal scholarship award amounts to entice them to enroll at the award-giving institution. One interesting aspect of this work is that data science applications in enrollment management remain very underexplored. This is true not only in terms of predicting enrollment but also scholarship disbursement, which are the two central aspects of this Chapter. Another interesting aspect of this Chapter is that it contains real-world results - we actually deployed the system we developed and saw how it fared. My involvement in this project came as the University was interested in having this work be conducted internally, after having it previously contracted out. Due to personnel shortages, I was approached with an opportunity to work on this problem, and it provided me with insight into university administration as well as added perspective on the role that data science can play in university management. In all, this work demonstrates how machine learning and numeric optimization can be used to better understand student's decisions to enroll at a particular institution.

2.2 Abstract

Effectively recruiting students and estimating student enrollment is critical to the success of any university. However, despite having an abundance of data and researchers at the forefront of data science, universities are not fully leveraging machine learning and data mining approaches to improve their enrollment management strategies. In this project, we

use data at a large, public university to increase their student enrollment. We do this by first predicting the enrollment of admitted first-year, first-time students using a suite of machine learning classifiers (AUROC = 0.85). We then use the results from these machine learning experiments in conjunction with genetic algorithms to optimize scholarship disbursement. We show the effectiveness of this approach using actual enrollment metrics. Our optimized model was expected to increase enrollment yield by 15.8% over previous disbursement strategies. After deploying the model and confirming student enrollment decisions, the university actually saw a 23.3% increase in enrollment yield. This resulted in millions of dollars in additional annual tuition revenue and a commitment by the university to employ the method in subsequent enrollment cycles. We see this as a successful case study of how educational institutions can more effectively leverage their data.

2.3 Introduction

Managing student enrollment is one of the core administrative tasks of any university. However, it is far from simple as universities aim to attract and retain the best students with limited resources [44, 83]. Enrollment management has wide-ranging implications on institutions' student body composition as well as their budgeting and finances, where a reliance on tuition income necessitates accurately forecasting student enrollments [81, 182]. One instrument that has continually been leveraged in the pursuit of enrollments and the associated tuition income is financial aid as receiving a financial aid award increases the likelihood of a student enrolling [102, 83]. While financial aid remains a powerful mechanism for institutions to reach their admissions and revenue targets, miscalculating projected student enrollments and mismanaging financial aid funds can have severe implications (such as rescinding over-committed offers[1])[6]. Furthermore, as institutions face tightening budgets and find their pricing policies continually under scrutiny, it remains imperative for them to optimize the resources they have by maximizing enrollments and the associated tuition revenue from fi-

[1]See https://bit.ly/2Scxqj6 as a recent example.

nancial aid programs [80, 86]. As such, accurately predicting enrollment and optimizing how student aid is disbursed is critical to enrollment management with financial implications that cascade across the entirety of an institution. In this work, we develop an approach to address this challenge, implemented it for a recent entering class, and found that it far outperformed previous strategies.

Predicting enrollment and optimizing the allocation of student aid requires data on student admissions and operational budgets. This data is stored in institutions' organizational databases or can be extracted from operational records. However, despite having this data on previous enrollments and finances, institutions are often slow to leverage it to gain actionable insights and improve institutional processes [153, 188, 26]. What's more, using data for insights in education is less prevalent at traditional campuses (i.e. schools where learning is primarily on-campus) and more common in online and computerized environments, which are more amenable to the collection and analysis of digitized data [125]. To this end, traditional universities remain "data-rich" but are "information-poor" in that they have the raw data needed to extract intelligible insights but are unable to do so due to infrastructure limitations and untrained personnel, among other reasons [159]. This results in the outsourcing of data-centric enrollment work (including developing scholarship disbursement and enrollment strategies) to full-service consulting firms, which do not disclose their proprietary approaches or how their results are evaluated [85]. The lack of motivation for consulting services to disseminate their work coupled with institutions trying to maintain competitive advantages in recruitment limits the extent of published research on how institutions can utilize data to improve recruitment processes. As a result, this dearth of literature provides little to demonstrate how data mining and machine learning can assist in the critical mission of enrollment management and in allocating financial aid.

In this project, we mine data from a large, public university in the United States (US) to optimize the disbursement of a merit-based scholarship for domestic non-resident students. We do this in two steps. We create a predictive model of student enrollment using historical student application data. We then use a genetic algorithm to optimize scholarship

disbursement to maximize student enrollment based on this predictive enrollment model. We conducted this work during the most recent admissions cycle of the university and the optimized awards were given to the latest entering class. After seeing improvement in student enrollment yield and an increase of millions of dollars in annual tuition revenue, the university incorporated our approach into their enrollment management process. We believe this project is a case study for other institutions seeking to similarly leverage institutional data for improving enrollment forecasting and financial aid allocation.

2.4 Relevant Work

The following discussion of relevant work is not exhaustive but is intended to give examples of relevant approaches with a focus on more recent work. While there is some work showing how to predict enrollment, there is very little showing how to allocate scholarships and hardly anything that ties the two together.

2.4.1 Predicting Enrollment

A few studies have employed machine learning and data mining techniques to predict university enrollment using non-neural approaches. DesJardins developed a logistic regression model using a dataset of approximately 14,400 students from an undisclosed tier I research university in the US. DesJardins' model gave an area under the receiver operating characteristic curve (AUROC) of 0.72 when predicting whether or not a student will enroll [58]. Similarly, Goenner and Paul used logistic regression to predict which of over 15,000 students at a US university would eventually enroll [71]. Their model gave an AUROC value of 0.87.

In addition to the above studies examining non-neural approaches for predicting enrollment, studies have also found that neural approaches fare very well for the same task and often perform better than non-neural approaches. For example, Walczak evaluated different neural network designs when predicting student enrollment at a US liberal arts college, stressing the problem as one of resource allocation [186]. Using a few thousand students, Walczak found that backpropagating neural networks fared best among those compared. Walczak and

Sicich later compared neural networks versus logistic regression to predict enrollment at two US universities [187], finding that neural networks performed better than logistic regression. Chang used logistic regression, decision trees, and neural networks to predict the enrollment of applicants at an undisclosed university, also finding that neural networks outperformed other models when judging by classification accuracy [40].

2.4.2 Scholarship Optimization

While there are some examples of works examining the use of machine learning in predicting enrollment, there is very little detailing scholarship disbursement strategies, especially ones leveraging machine learning and/or numerical optimization techniques. One example is the work of Alhassan and Lawal, who demonstrated the use of tree-based models for determining which students would be awarded scholarships in Nigeria [5]. Alhassan and Lawal describe the results as "effective" compared to approaches previously used but did not provide more on the success of their approach.

One work used machine learning to predict enrollment in conjunction with a numerical optimization technique to disburse scholarships. Sarafraz et al. used neural networks with genetic algorithms to optimize financial aid allocations and while our research is similar in spirit, there are a few notable differences [146]. Firstly, the scholarship fund optimized in this work is merit-based, meaning there are upper and lower bounds on scholarship awards that are specific to each student. This makes for a more difficult optimization task. We also examine alternative predictive models beyond just neural networks (such as ensemble approaches) and use a larger dataset in terms of both the number of observations and the number of features (over 72,000 observations vs 4,082; over 100 features vs 6). We also provide a comprehensive description of final model performance across multiple metrics and a detailed outline of how genetic algorithms can be used for aid disbursement, including a binning framework to drive the optimization task. Finally, we share real-world enrollment metrics after employing the scholarship optimization to demonstrate the effectiveness of our approach.

2.5 Methods

We present the methods for this work by first giving an overview of the setting; then, we describe the data and feature engineering; we then discuss how we predicted enrollment; finally, we discuss optimization constraints and the optimization process. The overall process for this work is shown in Figure 2.1. Due to the sensitive nature of the data and the fact that it contains personally identifiable information (i.e. student names, addresses, and high schools), we are unable to make it widely available. However, we present the methods below with as much transparency as we can so others can replicate the work.

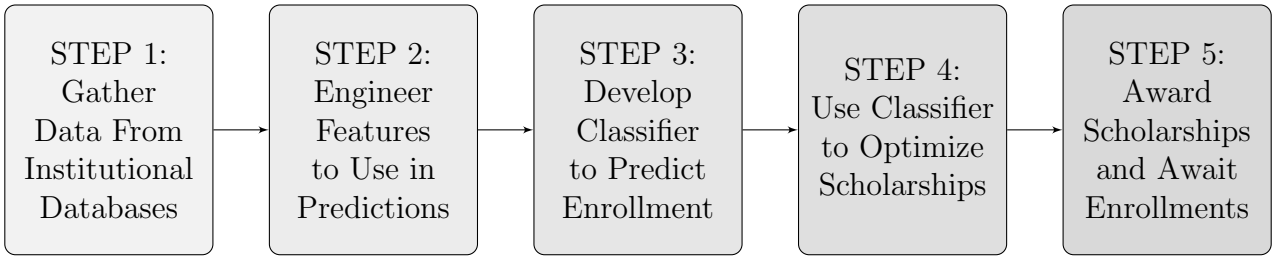

Figure 2.1: Process for optimizing scholarships, starting with data from University databases and ending with disbursements.

2.5.1 Setting

This scholarship optimization work was performed at a large, public US University (the University[2]) in early 2018. The scholarship fund examined was created to maintain the University's academic standards while maximizing the enrollment of first-time, first-year (freshmen) domestic non-resident (DNR) students by giving them financial incentive to attend the University. DNR students are students from the US who are not from the state in which the University is located. DNR students account for larger tuition charges than their resident counterparts so their enrollment is of high importance from a budgeting perspec-

[2]University administrative offices requested that the institution not be identified.

tive. Tens of millions of dollars in total are awarded annually to these students from the scholarship fund with millions eventually given to students who enroll.

The scholarship fund examined (DNR scholarships) was to be disbursed based on merit. As such, students with higher academic profiles, as defined later, were given equal or larger scholarships than those with lower academic profiles, regardless of financial need. Additionally, only freshmen DNR students who were accepted to the University were eligible for a DNR scholarship award. All admitted DNR students were automatically considered for a DNR scholarship and students did not need to apply for the scholarship.

In years prior, the disbursement strategies for the DNR scholarship were developed by external consulting services. For the last full application cycle (the 2018 entering class), the disbursement strategy was brought under the technical stewardship of the University. This is the application cycle for which we optimized scholarship disbursement. The models that were previously developed for the disbursement of the scholarship fund were proprietary to the consulting services and could not be leveraged. However, student application, enrollment, and scholarship data from prior years was available. When describing results, we compare the results from our approach to that developed by the consulting services. We cannot compare the approach detailed in this writing to a completely un-optimized approach or one that is randomized because the scholarship has never been disbursed in such a manner.

Award-receiving students concurrently learned of the amount of their scholarship and of their admittance to the University. However, not all applications were scored by admissions officers when the first awards were to be given. This was primarily due to the admissions review timeline at the University. We did not know of every admitted student at the time of optimization yet the scholarship awards were only to be given to admitted students. Thus, the last full application cycle's data could not be used directly in the optimizations. Instead, we used data from prior years to develop a fund allocation strategy and then apply this strategy to the last application cycle. This was with the expectation that applicants in the last application cycle were statistically similar to years prior across all the variables used in the modeling and we checked to ensure that this was in fact the case using individual t-tests.

2.5.2 Data

The data for this work consisted of information on all freshmen DNR applicants to the University from 2014-2017 with usable data. This totaled 72,589 students. The data was compiled from two major institutional sources: the students' admissions applications and their Free Application for Federal Student Aid (FAFSA) information. The FAFSA is an application prepared by incoming and current US college students to determine their eligibility for financial aid. Examples of data from students' admissions applications include their high school coursework, entrance exam scores, college GPA (if they had taken classes for credit), whether they were a first-generation college student, and their parents' educational attainment. These were all self-reported and verified by the University as needed. Data directly from and derived from student FAFSA filings included students' family income, their expected family contribution to college expenses (as calculated by the University), and loan amounts awarded to the student. About 66% of students had filled a FAFSA. Also included in the data were indicators of whether each student eventually enrolled at the University. Of the 72,589 students in the dataset, 5,081 enrolled (7.00% of all). Demographic variables such as gender and race were available but were not used as discussed in Section 2.6.1.

The data included tuition amounts students would pay on an annual basis, their financial aid grants and scholarships awarded (outside of DNR scholarship awards), and their DNR scholarship award amount. These variables were not included in any prediction or optimization model on their own. Instead, we created a "`reduced_tuition`" variable which was the annual tuition amount for the students less their total grants and scholarships (i.e. the other two variables summed). We used this variable as a single financial aid and tuition-related feature for the optimization process. This feature is not altered when developing the predictive classifier but is altered during the optimization task, during which the response of students to different award amounts are simulated.

2.5.3 Feature Engineering

Prior to prediction and optimization, we engineered features from existing variables. First, we either converted categorical variables to dummy variables or replaced them with a binary indicator variable. Then, we grouped students based on their FAFSA award amounts into 6 discrete bins (which were in line with University financial aid record-keeping), each of which was used as a categorical feature. We created binary indications of whether students attended each of the 10 most popular high schools for student applications and did the same for the 10 most popular states from which students applied. A binary indication was also created for a student athlete designation as each sport had its own application codes. In addition, we also created a separate binary indication for whether the student was transferring any credits from a college in high school program. Students also indicated their academic interests on their applications to the University. We pulled these from their applications and grouped them into 12 broader categories based primarily on the college/department they were at the University (e.g. "Engineering", "Humanities", "Health Sciences", etc.). We then created binary indications of whether a student was interested in each of the categories. Only students' first application to the University and the resulting admissions/enrollment decisions were included in the data. There were a total of 108 features.

Not all applicants filed a FAFSA form and we imputed missing FAFSA-related values. We performed this imputation by building a separate gradient-boosted regression tree model for each FAFSA-related feature using all features that were complete. We then used these regression models to predict the missing values. Only FAFSA-related values were missing and no other features needed to be imputed.

2.5.4 Predicting Enrollment

To predict enrollment, we first randomly divided the data using a 80-20 training-test split, with 57,359 students in the training set and 14,340 students in the test set. We did not re-balance the data with respect to classes. We scaled the training data by subtracting the

median of each feature and dividing by the feature's interquartile range. We subsequently scaled the test data using the scaling values from the training data. The binary outcome variable indicating whether the student enrolled at the University was not scaled.

After performing the training-test split, we trained 7 machine learning (ML) classifiers on the training set to predict enrollment. These classifiers were: a bagging tree ensemble (BC), gradient boosted trees (XGB), K-nearest neighbors (KNN), random forests (RF), regularized logistic regression (LR), support vector machines (SVM), and a neural network with 3 hidden layers (MLP). We tuned the hyperparameters for each of the classifiers using 5-fold cross validation on the training set. We report performance from all classifiers on the test set, which was not used to train the classifiers and only used to evaluate final performance. We used the classifier with the best performance to optimize scholarship disbursement.

2.5.5 Optimizing Scholarships

Genetic Algorithms

After developing a classifier to predict enrollment, we used the predictions from the classifier as an objective function in optimization. The aim of the optimization was to develop a strategy that maximized student enrollment from the DNR scholarships. In other words, the optimized approach disbursed scholarships in a manner that maximized the number of students who would enroll at the University from a pool of admitted students to the University. In this work, we used a genetic algorithm (GA) for optimization as GAs are known to work well with a well-defined measure to optimize (i.e. student enrollment) but not a well-defined, continuous, and/or differentiable objective function. GAs are also known to find near-optimal solutions quickly, which was essential when we wanted to rapidly outline different budgeting scenarios early in our modeling.

GAs are a class of evolutionary algorithms and are inspired by biological evolution. GAs generally involve iteratively starting with a population of chromosomes, undergoing selection across this population according to a measure of fitness, using genetic crossover and mutation

to produce offspring from the most fit individuals, and then using this offspring as the population for the next iteration [114]. The overall population fitness improves with each iteration and the GA eventually converges towards an optimal solution. In this work, we start with a population of award disbursement strategies whose "genetic material" (chromosomes) are a set of scholarship award values; the measure of fitness to assess these individuals is based on predicted enrollment after accounting for constraints; and the crossover and mutation functions used to create offspring are based on altering scholarship award values.

We used the data for the previous year's (2017) admitted class in the optimization of scholarship funds. In all, this was 9,479 students (N_{total}). In this sense, we used data from the year prior to optimize the disbursement for the most recent application year. We pared the data down to a single year's application cohort to avoid having to consider if any of the optimization constraints in Section 2.5.5 were being violated for each of the application years simultaneously.

Binning Students

We generated a set of possible scholarship awards that spanned S_{min} to a chosen maximum (S_{max}) in \$300 increments and included \$0. We did not determine S_{max} beforehand but instead set it such that the optimization procedure did not generate an output that included a S_{max} scholarship award. S_{min} was evenly divisible by \$300 and we generated possible scholarship awards in \$300 increments to satisfy constraint (4) from Section 2.5.5. In all, there were over 20 unique scholarship award values and only these award values were used in the optimizations.

Part of the difficulty of this particular optimization task lies in the fact that awards were to be given in a merit-based manner. As such, the scholarship award for any student is dependent on the awards of students with similar academic profiles. For example, if one was to rank all admitted students in the application pool based on a measure of merit, the minimum possible award given to a particular student would be determined by the award given to the student with the merit that is immediately lower. Similarly, the maximum

award for a particular student would be equal to the award given to the student with the merit that is immediately higher. As such, if optimizing on a per-student basis, altering the award for any given student to influence their enrollment decision could result in a cascade that subsequently effects every other student's award amount. This results in a very complex fitness landscape when optimizing scholarship awards individually.

As a solution to this issue of an optimization cascade, we first ranked and binned students based on academic merit such that all students in the same bin received the same scholarship award. To perform this binning, we sequentially ranked students based on 3 variables: their application academic score, their high school GPA, and their scores on college entrance exams, in that order. This ranking was students' "academic profile." Each student's application academic score was based on a holistic scoring of their academics and was the primary variable for determining their academic profile. We were provided this metric by the University admissions office. Ties between students having the same application academic score were broken by looking at their high school GPA; any remaining ties thereafter were broken using students' entrance exam scores. Once students were ranked, they were divided into 20 ventiles based on their academic profiles (i.e. students were grouped across every 5th percentile) with each ventile receiving the same scholarship award amount. Using ventiles allowed for us to have sufficient flexibility when exploring the fitness landscape during optimization while also not being so granular as to continually be caught in local extrema. Additionally, ventiles helped mitigate the effect of optimization cascades by giving identical awards to students with similar academic profiles. We refer to each of these ventiles as a "bin" and each bin served as the chromosomal building block for the GA. A single scholarship allocation strategy consisted of the scholarship awards across all 20 scholarship bins and is referred to as an "individual" henceforth when used in the context of the GA. Thus, each individual's genetic material can be thought of as being in the form of chromosomes composed of scholarship award bins. It should be noted that we used ventiles after examining the optimization results from other binning strategies (namely using 10, 15, and 25 bins) and finding them to give lower predicted enrollments. We did not, however, attempt to find

an optimal bin number beyond this but do intend to explore this in the future.

After binning students, we created a fitness function to evaluate the effect of altering the `reduced_tuition` variable on student enrollment. Specifically, this function took the genetic material of a scholarship individual (i.e. a set of scholarship awards for each bin) and then re-evaluated the `reduced_tuition` variable for each student based on their updated DNR scholarship award. As noted above, we created the `reduced_tuition` variable by taking the tuition due for a student and subtracting their total grants and scholarships; it was the only financial aid and tuition-related variable used in the predictive model. The function re-calculated each student's likelihood for enrollment based on the updated values for `reduced_tuition` using the predictive enrollment model. The final output for the fitness function was a calculation of the number of students predicted to enroll for a given scholarship individual, which we used as the fitness criterion for evaluating individuals.

Modeling Constraints

Several constraints were posed on the scholarship disbursement by University administrators. Due to University policy, exact values for awards and budgets will not be discussed. Some constraints on the disbursement strategy were as follow, where F represents funds in DNR scholarship offers, B represents funds in the DNR scholarship budget, N specifies a count of students, and S specifies a scholarship award amount:

1. The total amount spent on DNR scholarships (F_{spent}) cannot exceed a pre-determined amount (B_{spent}): $F_{\text{spent}} \leq B_{\text{spent}}$

2. The total amount offered to students in DNR scholarships regardless of whether they enroll (F_{offered}) cannot exceed a pre-determined amount (B_{offered}): $F_{\text{offered}} \leq B_{\text{offered}}$

3. The percentage of admitted students who are awarded scholarships ($N_{\%\text{awarded}}$) should be approximately equal to a pre-determined percentage ($N_{\%\text{target}}$): $N_{\%\text{awarded}} \approx N_{\%\text{target}}$

4. The award amounts must be divisible by \$300 to allow for round hundred-dollar splits across three academic terms.

5. There is a minimum value for a single scholarship award ($S_{\min}$) but no pre-determined maximum value.

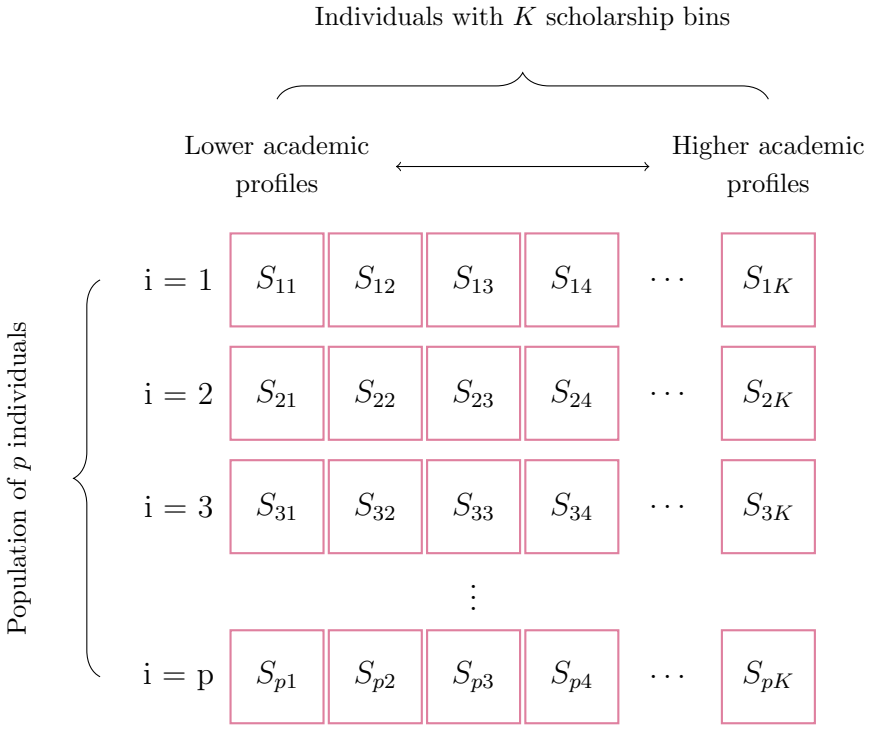

Figure 2.2: Genetic algorithm setup. Individuals (i) are scholarship allocation strategies of K scholarship bins (j). The population consists of p individuals. Each S_{ij} is a scholarship award value for the i^{th} individual and the j^{th} scholarship bin. The bins are sorted based on academic profile such that $S_{i1} \leq S_{i2} \leq S_{i3} ... \leq S_{iK}$ for any given i (but not necessarily across individuals). For this work, $K = 20$ and $p = 1000$.

The organization of the population, individuals, and bins for the GA optimization is shown in Figure 2.2. We generated an initial population of p individuals by randomly selecting K scholarship awards (one for each bin) from the set of possible scholarship awards and sorting for each individual. For this work, $p = 1000$ and $K = 20$. Each bin contained

the same number of students (N_{bin}), which was equal to $\frac{N_{\text{total}}}{K}$. All students in the same bin received the same award for a given individual; awards were not unique to each bin and could be duplicated across a given individual. N_{bin} multiplied by the scholarship award value for each bin equalled the funds awarded for that respective bin; the sum of these across all K scholarship bins for an individual was F_{offered} for that individual. The predicted number of enrollees for each scholarship bin multiplied by the award for that respective bin equalled the funds spent for that bin; the sum of these across all K scholarship bins for an individual was F_{spent} for that individual. The number of bins with non-zero award values divided by K was equal to $N_{\%\text{awarded}}$ for an individual.

We penalized each individual's fitness if the optimization constraints above were violated. We initialized a single penalty coefficient (σ) to 1.0 and then successively enforced each of the following squared penalties for a given scholarship individual:

- if too much was spent on scholarship awards:
  $F_{\text{spent}} > B_{\text{spent}} \rightarrow \sigma = \sigma * (\frac{B_{\text{spent}}}{F_{\text{spent}}})^2$

- if too much was offered in scholarship awards:
  $F_{\text{offered}} > B_{\text{offered}} \rightarrow \sigma = \sigma * (\frac{B_{\text{offered}}}{F_{\text{offered}}})^2$

- if too many students were awarded a scholarship:
  $N_{\%\text{awarded}} > N_{\%\text{target}} \rightarrow \sigma = \sigma * (\frac{N_{\%\text{target}}}{N_{\%\text{awarded}}})^2$

- if too few students were awarded a scholarship:
  $N_{\%\text{awarded}} < N_{\%\text{target}} \rightarrow \sigma = \sigma * (\frac{N_{\%\text{awarded}}}{N_{\%\text{target}}})^2$

Ultimately, we multiplied the output of the fitness function by the penalty coefficient to penalize constraint-violating individuals. If there were no constraints violated, the penalty coefficient was 1.0 and the fitness evaluation of the individual remained unchanged.

Optimization Process

The approach for the GA was as follows. We randomly generated the initial population of individuals as described above. We then calculated the fitness of each individual and took a subset of the most fit individuals (10%) as the basis for the next generation of the population. We employed genetic crossover to this subset of the population to generate offspring. We used two-point genetic crossover, wherein two points were randomly selected along chromosomes and the genetic material from one individual was swapped with that from another between the two points, much like a two-point crossover mutation in nature. In other words, for a pair of randomly selected individuals, we randomly selected two scholarship bins from ventiles 1 through 20 and all scholarship award values between the two bins from one individual were swapped with those from the other individual and vice versa.

After using crossover to refill the population, the offspring underwent mutation. We used three types of mutations: an increase mutation, a decrease mutation, and a swap mutation. For a mutation, we randomly selected an individual and then randomly selected a bin from this individual. The award for this bin was either increased to another possible award amount (increase mutation), decreased to another possible award amount (decrease mutation), or swapped for another randomly selected award amount (swap mutation). The probability of performing either an increase, decrease, or swap mutation were equal unless the scholarship award value equaled $S_{\min}$ or $S_{\max}$, in which case we eliminated the possibility of a decrease or an increase mutation, respectively. Once a particular mutation was selected for a given individual and bin, a single award value was randomly selected from all possible award values that satisfied the condition of the mutation and used in the mutation. After mutations, we re-sorted the awards across each individual to ensure students with higher academic profiles received larger awards. We kept the initial subset of the most fit individuals unchanged during crossover and mutation; instead, we altered replicas of these individuals to compare the most fit individuals from one generation to those from the next generation. The new generation of individuals then served as the population for the next algorithmic iteration. We

repeated this process for 20 generations of the population and used the most fit individual thereafter as the scholarship allocation strategy. The process for the GA is shown in Process 1.

Process 1: Genetic algorithm process for scholarship allocation
(parameters for this work are in parentheses)

1: Generate initial population ($p = 1000$ with $K = 20$ bins each)
2: Evaluate fitness of each individual (where fitness is
   enrollment count predicted by XGB classifier)
3: For each of G generations: ($G = 20$)
4: Keep subset of population with highest fitness (10% kept)
5: Use two-point crossover across individuals to fill population
6: Mutate random bins of random individuals
7: Evaluate fitness of each individual
8: Use individual with highest fitness after G generations

2.6 Results and Discussion

Using the methods described in Section 2.5, we developed a predictive classifier of student enrollment and used it in conjunction with a genetic algorithm that optimized the allocation of a scholarship fund. Ultimately, the university saw a 23.8% increase in enrollment yield after using our approach. This resulted in millions of dollars of additional annual tuition revenue. The following section presents these results in greater detail in the same order as the methods.

2.6.1 Predicting Enrollment

Previous studies have shown the effectiveness of ML in predicting enrollment. We examined seven different predictive classifiers for this task. We show the performance of these classifiers in terms of prediction accuracy, AUROC, and F1-score in Table 2.1. We used the same observations as a test set to compare performance across classifiers; for the test set, the majority class represented 92.8% of observations (i.e. 7.2% of students in the test set

Table 2.1: Classifier performance sorted by rank across all metrics. Names of classifiers are provided in Section 2.5.4.

	Model	Accuracy	AUC	F1-score
1.	XGB	93.10%	0.846	0.905
2.	RF	93.06%	0.848	0.901
3.	MLP	93.01%	0.845	0.902
4.	BC	93.05%	0.833	0.901
5.	LR	92.96%	0.805	0.900
6.	SVM	93.00%	0.780	0.900
7.	KNN	92.80%	0.793	0.893

eventually enrolled at the University). All classifiers performed similarly in terms of both accuracy and F1-score. Because of the large class imbalance, there were only modest gains in terms of accuracy over the majority class representation. Ensemble classifiers (RF, XGB, and BC) had the highest accuracies while KNN performed on par with the majority class representation (note: it was checked that the KNN model did not predict that all observations were of the majority class). The highest F1-score, meanwhile, was given by the XGB classifier, though it was not substantially higher than other classifiers.

We show ROC curves for the classifiers in Figure 2.3. The general shape of the ROC curves was similar across the classifiers but with meaningful variation in AUROC. Specifically, RF, XGB, and MLP tended to perform similarly in terms of AUROC and had the highest AUROC values. This is in line with previous work where neural networks tended to perform well when predicting enrollment, even without more complex architectures in this case. That said, the ensemble classifiers performed similarly well for the task at hand.

Demographic data was not used in the models. Including demographic variables in the prediction models would improve predictive performance to some degree, albeit at the expense of potential explicit discrimination with respect to recipient characteristics. As such, we decided to exclude demographic variables when building the classifiers. While doing so

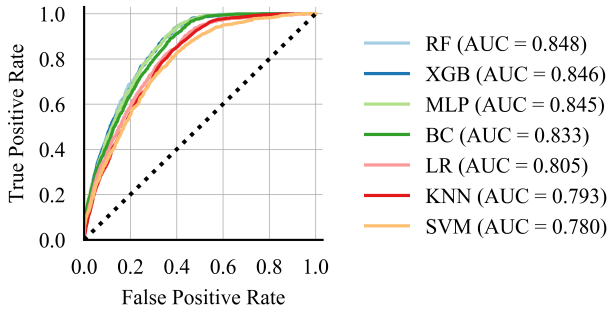

Figure 2.3: ROC curves for enrollment prediction

limits the degree of explicit discrimination, the possibility of implicit discrimination remains - particularly with respect to associations between demographics, income, geography, and academics. Checking and controlling potential demographic imbalances is beyond the scope of this particular work but was handled by stewards of the DNR scholarship fund after optimization.

We examined classifier performance across all metrics and decided to use XGB to optimize scholarship allocation. Prior to optimization, we calibrated the classification threshold for the prediction probability to the nearest one-hundredth such that the number of students predicted to enroll by the model was nearest to the actual enrollment count. By calibrating the threshold in this manner, we used a lower probability decision threshold (0.22) than the value of 0.5 that is typically used in binary classification. We understood that doing so came at the expense of an increased rate of false positives (Type I error) but it also allowed for the prediction counts to be closer to actual counts, which was necessary when discussing predictions with administrative stakeholders. We show the effects of this calibration in Figure 2.4, where the confusion matrix using the typical threshold of 0.5 is shown along with the confusion matrix using the calibrated threshold of 0.22.

Of note from the confusion matrices is how well students who did not enroll at the University could be identified. On the other hand, it was much more challenging to identify those who would enroll. This speaks to the selectivity of the University in that many of the

candidates who would not enroll were simply those who were not accepted to the University (students' acceptance to the University was not included as a feature during predictions). Concurrently, the difficulty with identifying students who will enroll aligns with the fact that these DNR students are applying to a university that is away from their respective homes and social bases. Also, those that are accepted to the University tend to be of higher academic standing, giving them more potential college choices. Thus, the general likelihood of a DNR student enrolling is difficult to determine when considering potential social factors and college options.

Lowering the classification threshold resulted in predicted enrollment counts in line with what was seen in the data, as shown in Table 2.2. Calibrating the classification threshold also allowed for a greater number of true positives while also balancing the number of false positives and false negatives. We also examined the effect of similarly calibrating the classification thresholds when using the other ML classifiers and determined that using XGB would still be viable for scholarship optimization.

2.6.2 Optimizing Scholarships

After we developed a model for predicting student enrollment, we used a GA to design a scholarship disbursement strategy. We used the GA in a setup with students grouped in

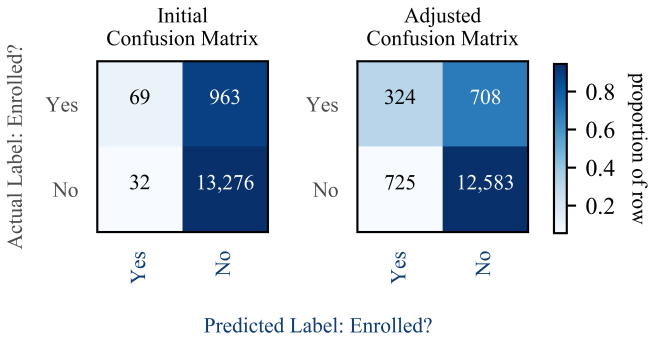

Figure 2.4: Confusion matrices for predicting enrollment using XGB and a classification threshold of 0.5 (left) and a calibrated classification threshold of 0.22 (right)

Table 2.2: Predicted enrollments after adjusting the classification threshold for test data and all data (training + test data).

	Test Data	All Data
Actual	1,032	5,081
Predicted	1,049	5,166

ventiles and each ventile receiving the same award amount. The genetic material (awards for each ventile) for individuals (allocation strategies) was altered for each iteration of the GA and then fitness was determined. Fitness was based on predicted enrollment after accounting for the violation of constraints. Due to the application review timeline at the University, we did not know which students of the most recent entering class (2018) would be admitted and used the prior year's application data (2017) to develop a disbursement strategy. Because the disbursement strategy relied on students being grouped into ventiles, we easily applied it to the most recent entering class after checking that the two classes were similar. Additionally, the binning strategy and the use of ventiles alleviated concerns about the size of the entering class as specific award amounts were disbursed to proportions of the entering class and not to a fixed count thereof.

We show fitness (predicted student enrollment) measures across the population of individuals for each generation of the GA in Figure 2.5. As expected, the maximum, mean, and

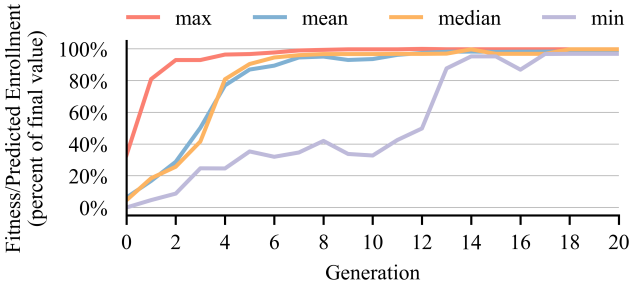

Figure 2.5: Fitness measures across generations of genetic algorithm. Fitness was equivalent to predicted enrollment.

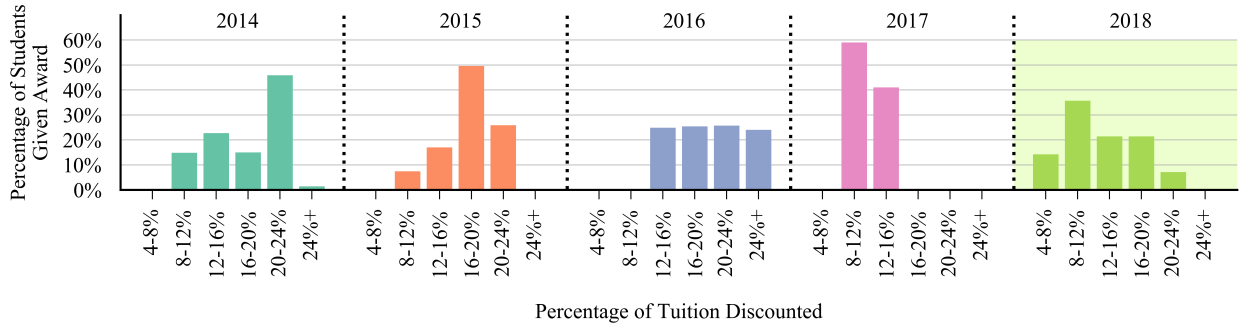

Figure 2.6: Historical scholarship allocations for the DNR scholarship. The highlighted year (2018) shows the optimized scholarship allocations from this work. Upper bounds for the bins are inclusive. Percentages are of award-receiving students only.

median values of fitness increase across generations, though these increases are much smaller for later generations. The minimum fitness values for the population follow a similar trend with some variation. All metrics eventually converge to the predicted enrollment, which is shown as a percentage. Monte Carlo simulations will be used in the future to outline a distribution of likely enrollment counts.

The exact award amounts for the DNR scholarship cannot be disclosed due to University policy. Additionally, the percentage of students receiving scholarship awards was not consistent across previous years. For example, in some years, 30% of accepted DNR students may receive a scholarship while in other years, 70% of accepted DNR students may receive a scholarship. Furthermore, tuition charges change annually at the University. Thus, in an attempt to provide a normalized measure for comparison across entering classes without disclosing exact award amounts, we compare award allocation strategies across time based on the discount on tuition. For example, a student receiving a $5,000 scholarship when tuition is $20,000 receives a 25% discount on tuition. We show previous allocations of the DNR scholarship to scholarship-receiving students as a discount on tuition in Figure 2.6. This discount on tuition factors in tuition cost for a full-time DNR student but not additional living or educational expenses (i.e. housing, food, books, etc). To further illustrate the use of discount on tuition, when looking at Figure 2.6, it can be seen that approximately 15% of

all scholarship-receiving students received an award that discounted their tuition by 8-12% in 2014 while in 2017, approximately 60% of students received a similar award. It is apparent from examining previous allocations that the manner in which the awards were historically allocated shifted greatly from year to year. As noted previously, these previous allocations were determined by a external consulting services and we could not leverage their underlying approach in this work.

We also show the scholarship allocation strategy for the 2018 entering class (for which the scholarship disbursement was optimized in this project) in Figure 2.6. This strategy tended to favor smaller scholarships, which aligns with the optimized allocation strategy that Sarafraz et al. reported [146]. In fact, scholarship stewards had initially placed a lower limit on the scholarship awards ($S_{\min}$) during modeling, which was equal to the lowest scholarship amount that had historically been awarded to students. This lower limit was between a 8-12% discount on tuition. After we discussed preliminary results of the optimization and the effectiveness of smaller awards with the scholarship stewards, it was determined that the lower limit on the awards would be changed to $\frac{S_{\min}}{2}$. Thus, the 2018 entering class had some scholarship awards that were lower than those received by previous entering classes. These lower awards discounted tuition by 4-8%. It is also noteworthy that the optimized disbursement strategy gave a distribution of awards that was right-skewed (with more of the awards being lower in value), in contrast to previous allocation strategies, which were predominantly left-skewed (with more of the awards being higher in value) or near uniform. This speaks to the idea that smaller scholarships awarded to students of lower merit may be more effective than larger scholarships are for those of higher merit (keeping in mind that students who received smaller awards were also of lower merit for this merit-based scholarship). This aligns with intuition that those with higher academic profiles have more college options and require additional recruitment, be it additional financial aid or in some other form.

After we developed the scholarship distribution strategy for the 2018 entering class, the University distributed scholarship awards to admitted DNR freshmen. We then waited as

Table 2.3: Historical, predicted, and actual yields after scholarship disbursement.

	Timeframe	Yield	% Increase
Historical	2014-2017	10-12%	N/A
Predicted	2018	13.9%	15.8%
Actual	2018	14.8%	23.3%

these students indicated their enrollment decisions a few months later. In recent years, the yield for DNR students at the University was about 10-12% with little/no increase, as verified by scholarship stewards, where "yield" refers to the percentage of admitted students who enrolled at the University. Historical yields were not based on an un-optimized or randomized scholarship allocation strategy but were the product of the scholarship allocations derived by external consulting services. Thus, because we were comparing the results from our approach to those from a previously optimized strategy (and not an un-optimized or random allocation strategy), we expected to see a modest improvement. Instead, we saw a much higher increase in yield. Table 2.3 shows the historical yields, the predicted yield based on our optimized approach, and the actual yield based on student enrollment for the 2018 entering class. When comparing to the upper bound on historical yield (12%), we anticipated that the scholarship optimizations would increase student yield by 15.8% (12% to 13.9%) based on the enrollment numbers we had seen during the optimizations (which was computed using XGB and the calibrated classification threshold). In reality, yield increased by 23.3%. This amounted to hundreds of additional students enrolling with each paying tens of thousands of dollars annually in tuition. There was also no discernible difference between the academic aptitude of students from the last application cycle and years prior. Overall, the net effect was an increase in millions of dollars in annual tuition revenue for the University. The University has since incorporated our approach into their enrollment modeling process for future disbursements of this scholarship fund. Of note is that the above yields are based on proportions of students that enrolled and the size of the entering class makes little difference

when comparing yields. The University also admitted roughly the same percentage of DNR students as years past and nearly all conditions during the application process were identical to previous entering classes. That said, the degree to which this increased yield can be causally attributed to the scholarship optimizations warrants further investigation. This may be in the form of A/B testing or some other controlled experiment.

2.7 Conclusions

In this work, we show how existing data at a university can be used to improve enrollment management. We combine machine learning with numerical optimization and use student application data at a public university to optimize a scholarship fund. We find that the optimized approach increased student enrollment and generated millions in tuition revenue. Our approach has been incorporated into the university's enrollment forecasting.

We show that ensemble classifiers can give strong performance when predicting enrollment and we use a binning strategy based on student merit to make the optimization task more tractable. This strategy eliminated the need for per-student optimizations, thereby limiting the complexity of the fitness landscape during optimization. After optimization, we see that smaller scholarship awards work better for maximizing enrollment. In all, the University had historically seen little/no increase in enrollment yield and we projected that our optimized approach would increase yield by 15.8%. In reality, enrollment yield increased by 23.3%.

Universities are at the forefront of training the next generation of data scientists and developing data-centric tools/techniques. However, they are far behind in applying data science to their own administrative data and processes. This project attempted to move them in this direction. Using a suite of machine learning tools, we were able to increase a university's revenue from a scholarship fund by millions of dollars. We think there are many similar opportunities to harness the power of data science in the realm of education administration, especially in resource allocation.

2.8 Publications/Presentations

- Lovenoor Aulck, Dev Nambi, and Jevin West. Optimizing scholarship allocation to improve enrollment. In *Conference on Information and Knowledge Management (CIKM)*, 2019 ***(conference paper - in prep)***

- Lovenoor Aulck, Dev Nambi, and Jevin West. Using machine learning and genetic algorithms to optimize scholarship allocation for student yield. *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2019 ***(poster)***

- Lovenoor Aulck, Dev Nambi, and Jevin West. Using machine learning and genetic algorithms to optimize scholarship allocation for student yield. *2019 KDD Workshop on DL4ED: Deep Learning in Education*, 2019 ***(workshop paper)***

Chapter 3

UNDERSTANDING THE EFFECT OF SCHOLARSHIPS ON STUDENT ENROLLMENT DECISIONS

3.1 PREAMBLE

This Chapter presents a continuation of the work from the previous Chapter. In the last Chapter, I examined whether we can predict enrollment and find an optimal scholarship disbursement strategy using machine learning and numerical optimization techniques. Here, I perform a post hoc analysis of the scholarship disbursement strategy that was developed to determine the impact scholarships have on students' enrollment. One particularly interesting aspect of this work is that because we were behind the proverbial curtain in terms of allocating scholarship awards, we had complete visibility into how they were disbursed and were able to leverage this in our analysis. What's more, this work is also unique in that it provides an analysis of a scholarship disbursement strategy that was optimized for a very specific purpose - maximizing enrollment. In this sense, the work in this Chapter doesn't examine whether the goal of the disbursement strategy was met (i.e. maximizing enrollment) but instead whether the strategy allocated scholarships in a manner that was statistically significant with respect to the eventual target outcome (i.e. whether a student enrolled).

3.2 Abstract

Financial aid plays a pivotal role in how universities manage funds and shape incoming student classes. However, the effects of financial aid policies are widely debated and often not generalizable across institutions. In this work, we examine financial aid in a very specific context: the effect of merit-based scholarships on the enrollment of domestic, non-resident students at a large, public US University. For our analysis, we rely on the fact that we had

complete visibility into the scholarship allocation process because we had devised the scholarship disbursement strategy. We leverage this understanding of the scholarship disbursement process to deploy a regression discontinuity design. The scholarship allocation strategy allowed us to analyze several sharp discontinuities concurrently. We find weak evidence that scholarships positively impact the enrollment of students who are of lower academic ability while most of our estimates resulted in no statistically significant effects. We also find that the magnitude of our effect estimates are sensitive to methodological decisions with respect to the regression discontinuity design. Namely, the choice of bandwidth and the manner in which it is optimized, be it with respect to the count of students or some measure of merit, impact results in a non-negligible manner.

3.3 Introduction

Universities and colleges have two primary means to shape the size and composition of their incoming student classes: their decision to admit students who have applied and their decision to give admitted students financial aid to entice them to enroll [184]. With respect to the former, acceptance policies and application reviews are largely dependent on the aims and mission of individual institutions. With respect to the latter, meanwhile, governments and institutions continue to make large investments in student financial aid as a means to increase enrollment and student success [92, 110]. For example, the US federal government spends upwards of $90 billion annually on financial aid support for students in the form of grants, subsidies, and tax expenditures [115].

Despite these large investments, the success of any institutional financial aid program hinges on deploying disbursement strategies that maximize the effectiveness of resource-constrained financial aid budgets. Also, while financial aid awards are a powerful mechanism for institutions to reach their admissions targets, mismanaging financial aid funds can have severe implications [6]. Thus, it is crucial that universities and colleges understand the impact that financial aid awards are having on students they are targeting so they can maximize the effectiveness of their limited financial aid budgets. However, literature on this

topic remains divided on whether financial aid increases college enrollment rates and alters college choice (e.g. [143, 31, 92, 63]), improves college performance (e.g. [49, 28, 46]), and increases college persistence and graduation (e.g. [160, 43, 158]). Furthermore, institution- and location-specific idiosyncrasies, such as how college price variation may be correlated with the number of community colleges in a given state, further muddle the degree to which the the effects of financial aid could be disentangled from other factors that may seem exogenous, but in actuality are not [92]. Thus, the institutional setting and context in which financial aid awards are disbursed matter as it is difficult to generalize findings.

In this work, we use a regression discontinuity design to estimate the effects of a merit-based financial aid scholarship on students' enrollment at a large, public University in Washington State. A regression discontinuity design is a quasi-experimental pre-post test design that examines the effects of a treatment (in this case, scholarship awards) on an outcome (in this case, enrollment) based on some cutoff with respect to treatment. The financial aid awards were given to first-time college students from outside Washington State. We perform our analysis after having first developed a scholarship disbursement approach using machine learning and numerical optimization approaches that was designed to maximize enrollment. Our intimate knowledge of the scholarship allocation process gave us complete visibility into the scholarship assignment process, which we leveraged in our study design. The disbursement strategy also provided many sharp discontinuities for us to analyze across the breadth of students' academic aptitude. What we find is some weak evidence that the scholarship awards positively effect enrollment for students of lower academic merit. However, we also find that these results are highly sensitive on methodological choices made in modeling, particularly with respect to how bandwidths are specified at/near the discontinuities.

3.3.1 Related Work

There is a large amount of literature that focuses on the effect of financial aid on student's enrollment decisions and academic performance. Many of these studies use quasi-experimental research designs to estimate these effects. In the interest of focusing on more methodolog-

ically relevant work, we focus this section on works that have used regression discontinuity designs.

Numerous studies have employed regression discontinuity designs to examine the effects of financial aid. The assignment variable used by these studies in their study designs has varied widely and has included: discontinuities with respect to students' standardized test scores (e.g. [31]), with respect to students high school GPA (e.g. [49]), discontinuities with respect to a calculated measure of academic merit (e.g. [184]), discontinuities with respect to federal grant eligibility and need (e.g. [143, 28]), and discontinuities along multiple assignment thresholds concurrently (e.g. [92]). Some examples of regression discontinuity-based analyses of the effect of financial aid on student outcomes are covered below.

Bruce and Carruthers used a regression discontinuity design when evaluating the effect of HOPE scholarships on college choice in the State of Tennessee [31]. The assignment variable used for the discontinuities was students' standardized test scores. They found that the merit awards examined had no impact on students enrolling in post-secondary education, though they did find some evidence that the awards increased the quality of institutions attended by some students. Rubin also employed a regression discontinuity design to examine the effect of Pell Grant eligibility on college enrollment using data from a large, longitudinal study [143]. Rubin found no effect on enrollment for students who varied in Pell Grant eligibility but had similar family incomes, acknowledging that the findings were in line with other work examining Pell Grants.

Blom and Canton used a regression discontinuity design to examine the impact that the SOFES student loan program implemented at private universities in Mexico had on academic performance [28]. Their work exploited discontinuities in the assignment of awards based on economic need and found that students who received the loan tended to perform better than those who did not. Kane used a regression discontinuity design to examine the effect of the CalGrant program on students' enrollment decisions [92]. The grant had multiple eligibility requirements across income, assets, and high school GPA. Kane reported an increase in college enrollment for grant-receiving students as well as an increase in private

college attendance amongst lower income students.

Goodman used a regression discontinuity design as well as a difference-in-difference approach to estimate the effect of the Adams Scholarship - a tuition waiver to Massachusetts State public colleges - on college choice [73]. Examining breaks in academic skill, Goodman found no difference in overall college enrollment rate but did find that the scholarships pushed more students to enroll in public colleges instead of private colleges. Similarly, Cohodes and Goodman analyzed the same scholarship across later student cohorts in a regression discontinuity design to estimate the effect of tuition waivers on college completion [43]. They found that scholarship use actually lowered college completion rates as students disregarded college quality when making their enrollment choices.

These papers demonstrate that the effect of financial aid on student outcomes has been studied using regression discontinuity designs. However, the results reported by these works vary widely and are highly dependent on institutional context, the outcome variable of interest, and the discontinuities examined. In this work, we are interested in whether receiving a financial aid award from a US institution increases the likelihood that a student from another US state will enroll at the award-giving institution. In this sense, our work most closely resembles that of Van Der Klaauw [184] and that of Curs and Harper [49].

Van Der Klaauw employed a regression discontinuity design to estimate the effect of financial aid offers on the enrollment of students at a particular college. In their experimental design, Van Der Klaauw exploited the manner in which scholarships were disbursed at this college, namely that the institution first ranked students based on a measure of academic merit and then awarded scholarships based on tiers across this ranking. The school's merit calculation was based on a weighted sum of students' standardized test scores and high school GPA.

Our work is similar to Van Der Klaauw's in two major ways: first, we also examine the effect of financial aid awards on students' likelihood to enroll at the award-giving institution and second, awards are also disbursed based on a ranking and tiering of students based on academic merit. However, there are also several differences. Chief among these is the fact

that in the case of Van Der Klaauw's work, the scholarship award amounts for students within each tier varied slightly based in part on partially subjective evaluations of students' application packages by college administrators. In our case, the scholarship awards are consistent across tiers, giving more sharp discontinuities for evaluation.

Curs and Harper, meanwhile, employed a regression discontinuity design to examine the effect that a merit-based scholarship awarded to out-of-state freshmen students by the University of Oregon had on first-year GPA. The University awarded the scholarship based on hard thresholds on the entering students' high school GPAs. Curs and Harper found that the scholarship had a significant positive effect on first-year GPA.

The work we present aligns with the work of Curs and Harper and the work of Van Der Klaauw in terms of context and methodology. Like Curs and Harper, we examine the effect of a merit-based scholarship awarded to out-of-state freshmen students at a large, public research university in the United States (US). In this sense, our work is similar in terms of context. However, while Curs and Harper examine the effect of a scholarship on GPA, we are interested in the effect of a scholarship on the enrollment of students at the award-giving institution, like Van Der Klaauw. Also, like Van Der Klaauw, we deploy a regression discontinuity framework that exploits the fact that students were ranked and tiered based on measures of academic merit. What remains unique about our work is that we perform a post hoc analysis of a numerically optimized scholarship disbursement strategy that was optimized to maximize enrollment of students at the award-giving institution.

3.4 Methods

The methods for this work are described in the following order. First we describe the institutional setting and the data used in this work. Then, we give an overview of the scholarship disbursement strategy. Thereafter, we describe our regression discontinuity design as well our calculations of academic ability, which was used as an alternative to ranking students based on merit. Finally, we give an overview of our cross validation approach for determining an optimal bandwidth for the regression discontinuity and a suitable model specification to

use in the analysis.

3.4.1 Setting

The scholarships analyzed in this work were distributed by a large, public US University (the University[1]) in early 2018. Award-receiving students concurrently learned of the amount of their scholarship and of their admittance to the University. They then had about two months to confirm their enrollment at the University. The scholarship fund examined was created to maintain the University's academic standards while maximizing the enrollment of first-time, first-year (freshmen) domestic non-resident (DNR) students by giving them financial incentive to attend the University. DNR students are students from the US who are not from the state in which the University is located. DNR students account for larger tuition charges than their resident counterparts so their enrollment is of high importance from a budgeting perspective. Tens of millions of dollars in total are awarded annually to these students from the scholarship fund with millions eventually given to students who enroll. Students are given the award in equal installments across four years as a tuition waiver.

3.4.2 Data

The data for this work consisted of the application information for all DNR students who applied to the University and were admitted as part of the Fall 2018 entering class for whom data was available. In total, this was 8,177 students. The data included information compiled from two major institutional sources: the students' admissions applications and their Free Application for Federal Student Aid (FAFSA). The FAFSA is an application prepared by incoming and current US college students to determine their eligibility for financial aid. Examples of data from students' admissions applications include their high school coursework, entrance exam scores, college GPA (if they had taken classes for credit), whether they were a first-generation college student, and their parents' educational attainment. These

[1]University administrative offices requested that the institution not be identified.

were all self-reported and verified by the University as needed. Data pulled directly from and derived from student FAFSA filings included students' family income, their expected family contribution to college expenses (as calculated by the University), and loan amounts awarded to the student. About 66% of students had filled a FAFSA.

Also included in the data were indicators of whether each student eventually enrolled at the University. Of the 8,177 students in the dataset, 1,030 enrolled (12.6% of all). All students in the dataset were admitted to the University. Their college options and choices beyond their decision on whether to enroll at the University were not known.

3.4.3 Scholarship Distributions

The students were given a scholarship award of one of six different amounts. Due to University policy, we cannot disclose the exact amount of the scholarships that were awarded to students. The awards were as follow (where S_{max} is the maximum scholarship award amount): \$0, 17.9% of S_{max}, 35.7% of S_{max}, 64.3% of S_{max}, 85.7% of S_{max}, and S_{max}. Henceforth, we refer to these scholarship award amounts as (listed in the same order as above): $S0$ (\$0), $S1$, $S2$, $S3$, $S4$, and $S5$ (S_{max}). Students were awarded these scholarships in accordance with the machine learning and optimization process outlined in work we have previously documented [11]. In particular, we used data from previous students' applications to the University to first develop a machine learning classifier that predicts enrollment. Then, we used this classifier in conjunction with a genetic algorithm to optimize the disbursement of scholarship awards with the goal of maximizing student enrollment at the University.

The scholarships were merit-based awards. Financial need or demographics were not considered when allocating scholarships; only students' academic merit was considered when allocating awards. Students did not need to apply for the scholarships separately as all DNR students who applied to the University were automatically eligible for an award. Students' merit was determined by first ranking students according to their "academic profile" and then splitting the students into ventiles. Every student in the same ventile received the same scholarship award. Academic profile was determined by sequentially ranking students based

on 3 variables: their application academic score, their high school GPA, and their scores on college entrance exams, in that order. These three variables comprise the entirety of the information used to determine academic profile and student merit. No other information was used to differentiate award amounts among students. Each student's application academic score was based on a holistic scoring of their academics and was the primary variable for determining their academic profile. We were provided this metric by the University admissions office. Ties between students having the same application academic score were broken by looking at their high school GPA; any remaining ties thereafter were broken using students' entrance exam scores. Once students were ranked, they were divided into 20 ventiles based on their academic profiles (i.e. students were grouped across every 5th percentile) with each ventile receiving the same scholarship award amount. The award amounts were then optimized using machine learning and a genetic algorithm.

The final distribution of scholarship awards was determined by the scholarship stewards at the University and is shown in Table 3.1. This distribution of awards did not precisely follow the ventiles outlined by the scholarship optimization. That said, the students were still sorted along academic profile and those with higher academic profiles received scholarship awards that were equal to or larger than those with lower academic profiles. As such, the students with the lowest academic profiles among those who received a particular scholarship award were academically similar to students with the highest academic profiles among the students who received the next lowest scholarship award. Similarly, the students with the highest academic profiles among those receiving a particular scholarship award were academically similar to students with the lowest academic profiles among the students who received the next highest scholarship award.

3.4.4 Regression Discontinuity Design

To understand the effect of the scholarships on student enrollment, we employ a regression discontinuity design (RDD), which was first introduced by Thistlethwaite and Campbell in 1960 [173]. An RDD is a quasi-experimental study design that can be used to estimate the

Table 3.1: Scholarship Awards.

Award	Amount (% of S_{max})	Count of Students	% of All Students	Enrollment Rate
$S0$	0	1378	16.85	17.05%
$S1$	17.86	900	11.01	14.44%
$S2$	35.71	4025	49.22	12.40%
$S3$	64.28	916	11.20	10.70%
$S4$	85.71	504	6.16	8.33%
$S5$	100	454	5.55	5.73%

causal effects of policy while controlling for unobserved individual heterogeneity, such as that which may be correlated with scholarship awards [49, 101]. In an RDD, the causal effects of interventions are estimated by assigning some cutoff threshold at which an intervention is administered that is outside of the subjects' control. The average effect of treatment is then estimated by treating the data similar to a locally randomized experiment near the threshold and comparing observations on both sides of the threshold [101]. This inherently relies on the assumption that observations on either side of a discontinuity do not significantly differ in terms of other covariates and we used independent t-tests and chi-squared tests to ensure that this was in fact the case.

Here, we leverage the fact that the students were assigned a "treatment" (i.e. a scholarship award) while being ranked/sorted along a continuous assignment variable (i.e. "academic profile") without being able to manipulate it in any way after their application to the University. Because we were directly involved in developing the disbursement strategy of these scholarship awards, we have visibility into the allocation process and the manner in which students were given scholarships. We are also assured that students were not able to deliberately change their position along the assignment variable with the goal of receiving a different award. Additionally, we are also assured that students were not aware of the cutoffs that were used to determine treatment, thereby eliminating the possibility of students at-

tempting to improve/change their score when they just barely miss a cutoff threshold [161]. When looking at the scholarship disbursements, as shown in Table 3.1, there are 5 distinct jumps in award amounts and, thus, 5 distinct thresholds (i.e. discontinuities) that can be analyzed. For reasons described in Section 3.4.5, we ultimately used 3 of these were in our analysis.

3.4.5 Calculating Ability

When distributing scholarships, we ranked/sorted students relative to each other. However, this relative sorting of students can pose some difficulties when comparing students along the assignment variable. For example, two students who may be identical academically besides a slight difference in high school GPA (e.g. GPAs of 3.2 vs 3.3) may be ranked sequentially. At the same time, two students who have vastly different high school GPAs (e.g. GPAs of 3.4 vs 4.0) may also be ranked sequentially, provided there is no other student that slots between them when calculating academic profile. Thus, though academic profile is a continuous assignment variable, the difference between the academic ability of a student and those who are ranked next to them will vary based on how many similar students are in the applicant pool. The importance of this is apparent when one considers that the thresholds for scholarship awards were not devised according to academic profile but based on student ventiles/percentiles. To counter this, we developed a metric for our analysis that we call "ability" and we used it to get an alternate comparison of student based on academics.

To calculate ability, we leveraged the fact that students' academic profile consisted of a sequential ranking of three bounded discrete variables (i.e. application academic score, high school GPA, and entrance exam scores). First, we normalize each of these variables by subtracting their respective minimum and dividing by the difference between their respective maximum and minimum (such that their values were between 0 and 1). Then, we set each successive variable used in the ranking to be weighted slightly less than a one unit increase in the variable that was previously used in the sorting. For example, there were 5 distinct values for application academic score after scaling (0, 0.25, 0.5, 0.75, and 1.0) and because it

was the first variable used in the sorting, we gave it a weight of 1.0. There were 80 possible values for high school GPA, which originally ranged between 3.2 and 4.0 before scaling and was in increments of 0.0125 after scaling. We assigned a weight of 0.2499 to high school GPA, corresponding to a value that is slightly less than a unit increase in scaled application academic score. Similarly, we assigned a weight of 0.0124 to students' scaled test scores, corresponding to a value that is slightly less than a unit increase in scaled high school GPA. Thus, the calculation for ability was as follows:

$$\begin{aligned} \text{ability} = (1.0 \ \times \ & \textit{scaled academic score}) + \\ & (0.2499 \ \times \ \textit{scaled high school GPA}) + \\ & (0.0124 \ \times \ \textit{scaled test score}) \end{aligned} \tag{3.1}$$

After calculating ability for each student, we normalized the measure by subtracting the minimum value and dividing by the difference between the maximum and minimum values. This normalized ability measure is what we use henceforth, with the student with the highest ability measure given a score of 1.0 and the student with the lowest ability measure given a score of 0.0. Figure 3.1 shows the ability values plotted against the order (i.e. ranking) of students based on academic profile. Also shown in the Figure are the points where there are discontinuities in scholarship disbursement and the associated awards. Though ability provides a montonically increasing measure that is positively correlated with order, there are some distinctions of note. Firstly, there are clear breaks in the values for ability where there is a jump in application academic score (i.e. the first variable used in sorting students). This is due to the fact that application academic score had the highest weight of all variables in determining ability. It also emphasizes the idea that though two students may be ranked closely when sorting and ordering in terms of academic profile, they may be very different in terms of academic performance in the eyes of the University, as the application academic score is a holistic score of student academics assigned by the University. What's more, the

size of this jump varies as one moves towards students with higher ability. Examining Eq 3.1, the degree to which there is a jump in ability is dependent on academic score firstly but also dependent on high school GPA. A distribution of GPA values that is predominantly left-skewed will result in ability values with a much more pronounced jump. Thus, the jump in ability between students with an ability near 0.25 is much less pronounced than the jump between students with an ability near 0.75 as GPAs are more left-skewed for these higher ability students.

We chose to exclude the two highest scholarship award amounts ($S4$ and $S5$) from our analysis. This was because their associated discontinuities with respect to scholarships were very near locations of jumps in application academic score, as can be seen in Figure 3.1. In part, these jumps indicate regions where ability is discontinuous, thereby threatening the internal validity of the experiment [88]. Elaborating further, when analyzing either of these discontinuities, there were few students we could have included in the model before we would have to include students with different application academic scores. There is then no assurance that these students would be academically similar to the other students included in the model. This would, in turn, limit our ability to experiment with different bandwidths, as detailed in Section 3.4.6. Ultimately, we looked at only three discontinuities - the discontinuities between $S0$ and $S1$, between $S1$ and $S2$, and between $S2$ and $S3$. Apparent from Figure 3.1 is that each of these discontinuities had plenty of support in terms of the number of students on either side of the discontinuity before reaching a jump in academic ability.

We show the enrollment proportions for students across the discontinuities included in the analysis in Figure 3.2. Of note is the generally negative correlation between enrollment proportion and ability. This is also shown across the enrollment rate for each of the scholarship awards as listed in Table 3.1. This is not surprising - students with better academic performance will have more college options and are thus less likely to enroll at any one university. Figure 3.2 also shows evidence of differences in general trends in enrollment across each of the discontinuities.

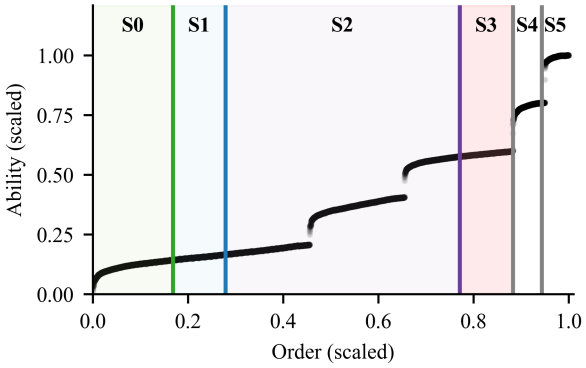

Figure 3.1: Ability of students across their order (ranking) based on academic profile. Shaded regions indicate scholarship award amounts and vertical lines indicate discontinuities across these awards.

3.4.6 Cross Validation

Regression discontinuity estimates are sensitive to the number of observations selected on either side of a discontinuity. In the context of this work, if too few students are included in the analysis, there is less data and statistical power to build a model that accurately captures any effect near the discontinuity. If too many students are included in the analysis, the model will be biased towards global trends rather than the effect at/near the discontinuity. This second fact is particularly important when assessing Figure 3.2 as there is an overall downward trend in enrollment across ability yet the localized difference in enrollment rates across a discontinuity may not follow the same trend. In this sense, smaller bandwidths reduce the bias of an estimate while larger bandwidths reduce its variance [151, 161]. In addition, regression discontinuity estimates are also sensitive to the functional form (in terms of polynomial order) used in the analysis. A misspecification of the functional form can lead to a bias in the estimate of the treatment effect [101].

To find an optimal combination of window size (i.e. bandwidth) and functional form to use in our analysis, we used 5-fold cross validation. The bandwidth and functional form (i.e. polynomial order) were both altered during cross validation. This process is in line with that

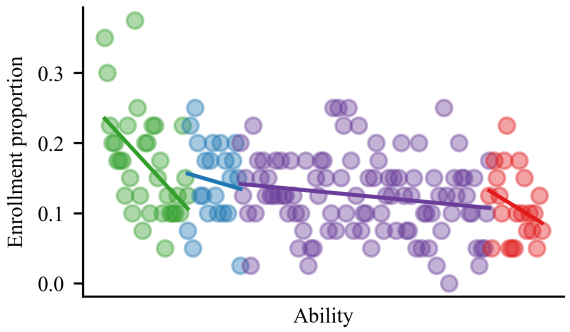

Figure 3.2: Enrollment proportions for students across ability. Colors correspond to scholarship awards in Figure 3.1. Each point shows the average of 40 students. Trendlines were fit independently for each scholarship award using a first order polynomial.

proposed by Lee and Lemieux [101] and we describe it in greater detail below.

Bandwidth Estimation

To determine an appropriate bandwidth for the analysis, we used two different approaches - one based on counts of students (and, accordingly, based on their academic ranking as used to disburse scholarships) and another based on ability, as calculated above. When examining bandwidths using counts of students, their ability was not taken into account. Rather, a fixed and equal number of students were selected on both sides of the discontinuity and included in the modeling. When examining bandwidths using the ability of students, only students within a specified range of ability values from the discontinuity threshold were included in the modeling. In both cases, we started by first determining a maximum bandwidth size for both sides of the discontinuity. When examining counts of students, the smaller of the counts of students who received the same scholarship on either side of the discontinuity was used as a maximum bandwidth size. When examining ability, the smaller of the maximum differences in ability amongst students who received the same award on either side of the discontinuity was used as a maximum bandwidth size.

During cross validation, we iterated across proportions of the maximum bandwidth size.

This meant keeping either counts of students or ability constant while iterating across the other. For counts of students, we included an equal number of students on either side of the discontinuity, regardless of ability; for ability, we kept the range of ability on either side of the discontinuity equal, regardless of the number of students included. During cross validation, we simply altered either the count or ability as a proportion of the maximum bandwidth.

Model Specification

To determine the model specification for the analysis, we began with the following functional form:

$$Y = \alpha + X\beta + XD\gamma + D\tau + \epsilon \tag{3.2}$$

where Y is enrollment, X is ability, D is a binary indication of whether $X >$ threshold (i.e. whether the student received the higher of the scholarship awards across a discontinuity), and XD is an interaction term that accounts for the varying effect of ability on enrollment. Further, α represents the average enrollment for students where all other independent variables are 0 (i.e. the y-intercept) and β and γ are the coefficients for ability and the interaction term, respectively. ϵ is the catch-all error term for the model and τ represents the estimate of the effect of a scholarship on enrollment. Of note is that in this functional form, the treatment is represented in a binary manner and indicates whether a student received the higher of the two scholarships across a discontinuity. We do not take the value and amount of the scholarships into account. In addition, we also tried the following functional form:

$$Y = \alpha + X\beta + D\tau + \epsilon \tag{3.3}$$

where Equation 3.2 and 3.3 are identical apart from the inclusion of the interaction term. We ultimately include both sets of results (including and excluding the interaction term) as its inclusion is primarily based on the modeling assumptions made - if one is to believe that there is no differential effect of ability on enrollment within the window examined during the

RDD analysis, the interaction term is unnecessary.

During cross validation, we iterated across different model specifications by changing the polynomial order of Equation 3.3. As such, during cross validation, the model specification took the following form:

$$\begin{aligned} Y = \alpha &+ X\beta_1 + X^2\beta_2 + ... + X^n\beta_n \\ &+ XD\gamma_1 + X^2D\gamma_2 + ... + X^nD\gamma_n \\ &+ D\tau + \epsilon \end{aligned} \tag{3.4}$$

where n represents the polynomial order. Of note is that for the above model specification, τ still represents the estimate of the effect of a scholarship on enrollment. Also, because D is binary and positive, its values remain the same even for higher order polynomial forms (i.e. $D = D^2 = D^3 = ... = D^n$).

Cross Validation Process

We used 5-fold cross validation to find an optimal bandwidth (in terms of both count and in terms of ability, separately) and a suitable polynomial order of our model based on Equation 3.4. We performed the cross validation separately for each of the 3 discontinuities analyzed - each had their own optimal bandwidth in terms of count and in terms of ability, along with an optimal polynomial order associated with each bandwidth. We simultaneously altered the bandwidth and functional form during cross validation. When altering bandwidth, we iterated across different proportions of a maximum bandwidth, as described in Section 3.4.6. When altering polynomial order, we iterated across different values of n in Equation 3.4. During cross validation, we predicted student enrollment (i.e. Y in Equation 3.4) and compared polynomial order/bandwidth combinations based on their area under the receiver operating curve (AUROC). This is a departure from the recommendations of both Lee and Lemieux [101] as well as Skovron and Titiunik [161], who suggest to use a bandwidth that

minimizes mean squared error (MSE). We opted to use AUROC instead of MSE for two primary reasons: 1) the outcome of interest in our case (student enrollment) is not continuous but binary and 2) the outcome is highly skewed in favor of non-enrollments (87.4% non-enrollments vs 12.6% enrollments). Thus, because we predicted a binary outcome variable during the cross validation process, we deployed the model specification from Equation 3.4 as a logistic regression. Specifically, the form of this logistic regression was as follows:

$$p(x) = \frac{1}{1 + e^{-(RHS\ Eq(4))}} \tag{3.5}$$

where $(RHS\ Eq(4))$ represents the right-hand side of Equation 3.4. We then opted to use a measure of performance that would not be inflated by having highly disproportionate classes and chose AUROC. The optimal values for each discontinuity were the combination of polynomial order and bandwidth size that maximized the AUROC.

3.4.7 Calculations of Effect Estimates

We used the optimal bandwidth/polynomial order combinations for each of the discontinuities from cross validation to produce our final estimates of the effect of receiving a higher scholarship on enrollment likelihood. These estimates are the values of τ from Equation 3.4. We also tested the robustness of these measurements by building a regression model of the appropriate polynomial order and across the same bandwidth that also included all covariates available from the student applications. We then compared the results from these models with the original results.

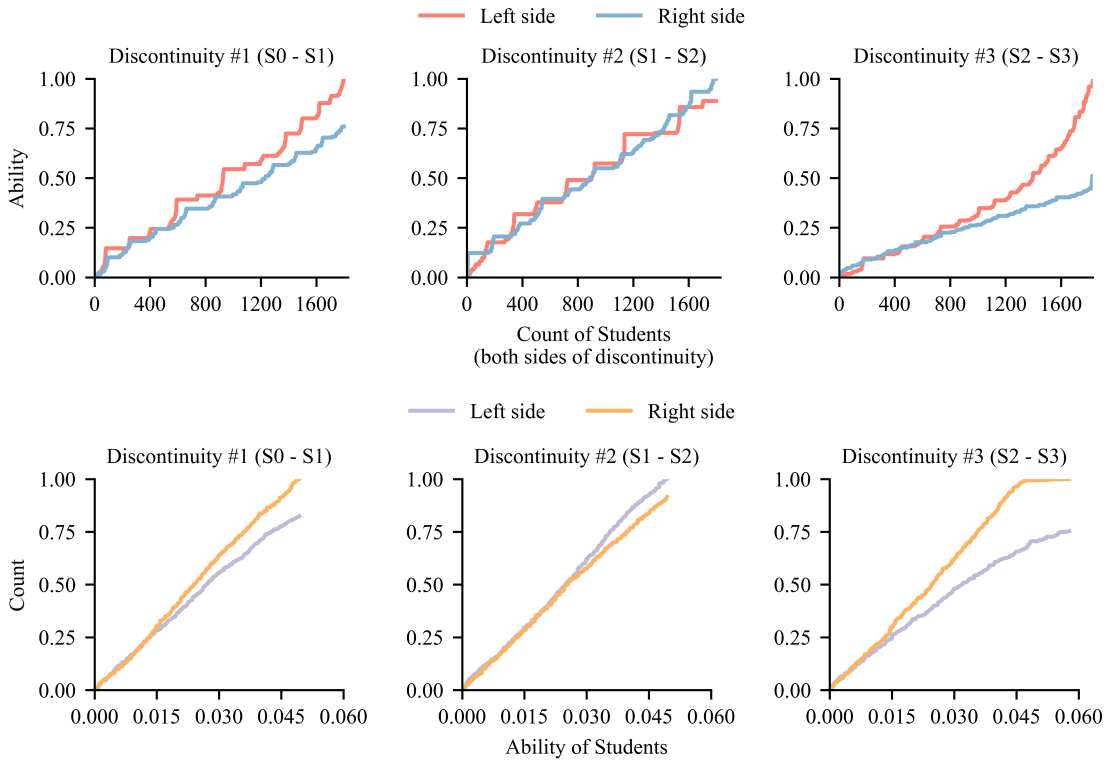

Figure 3.3: Ability of students as a function of count of students when optimizing bandwidths based on counts (top). Count of students as a function of ability of students when optimizing bandwidths based on ability (bottom). Y-axis values have been scaled relative to the maximum value for each subplot. X-axis values are across the maximum values in terms of count (number of students) and ability (%age of maximum).

3.5 Results and Discussion

3.5.1 Cross Validation Results

Asymmetries using counts and ability

We show the results from the cross validation process to optimize bandwidth and polynomial order in Figure 3.3. We created the top row of Figure 3.3 by iteratively adding a single student to each side of a discontinuity and then noting the change in the relative value of ability across each side of the respective discontinuity. From Table 3.1, it can be seen that the

maximum number of students that could be included during this process was similar (either 900 or 916) across each of the three discontinuities. The ability values on the left and right side, however, varied widely across the first and third discontinuities ($S0 - S1$ and $S2 - S3$, respectively) but were relatively similar for the second discontinuity ($S1 - S2$). Across the third discontinuity in particular, there was a much broader range of ability for students as one moved farther from the discontinuity on the left side which was not mirrored on the right side (as demonstrated by the convexity of the values on the left side).

The top half of Figure 3.3 highlights several things of note. Firstly, though the ability ranges of students can vary widely as one increases the bandwidth in terms of the number of students, this is not true when examining smaller bandwidth sizes. This is particularly noteworthy because an RDD estimates effects near the discontinuity. Thus, we can be assured that the students within smaller bandwidths are, in fact, similar in terms of ability. Secondly, using a bandwidth based on counts results in some jaggedness with respect to increases in the values of ability. This is not surprising, as students with similar ability levels may be clustered together. This also highlights some of the difficulty when using a bandwidth based on counts of students as the ranking may not be reflective of how similar/dissimilar students are in terms of their academic profile.

The bottom half of Figure 3.3 shows how the relative count of students changes as one increases the bandwidth in terms of ability. The general shapes of these plots are similar to the top half of Figure 3.3 but with swapped x-y axes. We created these plots by iteratively increasing the maximum ability range from the discontinuity. Compared to the top half of Figure 3.3, the lines along the bottom half of the Figure are far less jagged. In all, the overall trends still manifest for the plots - both the left and right side are fairly consistent in terms of size when the bandwidth is small but differ significantly for larger bandwidths.

Optimal bandwidths and polynomial order

We show the optimal bandwidth sizes in terms of both count and ability in Table 3.2. The values of ability are shown as a percentile of the total range of ability for all students. In this

Table 3.2: Cross validation results. For instances where two values are listed, the first value represents the left side of the discontinuity (i.e. students with lower awards) and the second value represents the right side of the discontinuity (i.e. students with higher awards).

Disc.	By Count		By Ability	
	# of students	Ability range (% of max)	# of students	Ability range (% of max)
$S0 - S1$	252	0.544 & 0.477	29 & 35	0.151
$S1 - S2$	396	0.904 & 0.477	152 & 143	0.822
$S2 - S3$	128	0.080 & 0.247	39 & 36	0.149

case, a value of 0.151 represents 0.151% of the maximum ability range across all students who were eligible for a scholarship. In Table 3.2, the range of ability represents the difference in ability between the student closest to the discontinuity and farthest from the discontinuity across a particular side. Of particular note is that the optimal bandwidth sizes when using ability tended to be smaller in terms of both the number of students included and the range of ability across students than those using counts.

The bandwidths also tended to be only slightly asymmetric when optimizing for ability. In this case, the discontinuity between $S0$ and $S1$ had the largest difference in terms of the number of students with the right side being 121% as large as the left side. However, these differences were much more pronounced when optimizing in terms of count. In this case, the discontinuity between $S1$ and $S2$ saw the left side have an ability range that was 190% of that of the right side. For the discontinuity between $S2$ and $S3$, the difference was more pronounced with the right side having an ability range that was 309% of that of the left side. This again highlights the difficulty in relying on counts of students after ranking as the sole measure to determine an optimal bandwidth.

When optimizing for polynomial order, we found that a first order polynomial was the best model specification across each of the discontinuities, regardless of whether we were looking at count or ability. Higher order polynomials did not significantly improve the cross-

validation results and we followed the recommendation of Skovron and Titiunik in choosing the lowest odd-numbered polynomial order possible [161]. When estimating effects, we did try to see the extent to which higher order polynomials altered our estimates and did not find any meaningful variation. Thus, all estimates shown use a first order polynomial fit (i.e. $n = 1$ in Equation 3.4).

3.5.2 Estimates of Effects

We show our our estimates of the effect size of receiving a scholarship on DNR enrollment at the awarding institution in Tables 3.3 and 3.4. We also present these results in graphical form in Figure 3.4, wherein the top bar for each discontinuity shows the estimates for bandwidths optimized in terms of ability and the bottom bar shows the estimates for bandwidths optimized in terms of counts of students. It should be noted that when we included interaction terms, the interaction term coefficient was not statistically significant for any of the models, across all discontinuities and bandwidth optimizations (the minimum p-value across all models was 0.23). This inability to reject the null hypothesis with respect to the coefficient of the interaction term being non-zero implies that there may not be any varying effect on enrollment across ability within the bandwidths examined.

In general, the effect estimates were either slightly positive or near-zero for each of the discontinuities. However, there was often a large difference in the estimates of the effect size when comparing count-based and ability-based bandwidths. This speaks to the importance of bandwidth selection in an RDD design as different bandwidths optimization schema could result in different results. In most cases, the general trends for the effect estimates did not vary with respect to whether they were positive or negative but indications of statistical significance ($p < 0.05$) did vary.

We see that the estimates when using bandwidths optimized by count are practically identical regardless of whether interaction terms are included or not. The estimates when using bandwidths optimized by ability, meanwhile, are slightly more positive for each of the discontinuities. The 95% confidence intervals for estimates without interaction terms are

Table 3.3: Estimates of effects of scholarship receipt on student enrollment without including interaction terms.

Disc.	Bandwidth	Estimate	95% CI	p-value
$S0 - S1$	By Count	0.077	(-0.077, 0.230)	0.33
	By Ability	0.267	(0.010, 0.523)	0.04*
$S1 - S2$	By Count	0.141	(0.019, 0.263)	0.02*
	By Ability	0.154	(0.041, 0.268)	0.01*
$S2 - S3$	By Count	0.009	(-0.173, 0.192)	0.92
	By Ability	0.002	(-0.213, 0.217)	0.98

* $p < 0.05$

also slightly more narrow than the 95% confidence intervals for estimates that included interaction terms. This slight downstream shift in terms of point estimates as well as narrower confidence bounds result in the effect estimates across the first two discontinuities when using bandwidths optimized by ability to be statistically significant for estimates not including interaction terms and not statistically significant when including interaction terms. Meanwhile, when examining the results based on the functional form with interaction terms, in only one case was the estimate statistically significant. This was when examining the discontinuity between $S1$ and $S2$ using a bandwidth optimized based on counts. In this case, the effect estimate was relatively small but positive. It should be noted that the same bandwidths were used to estimate effects, regardless of whether interaction terms were included.

In general, point estimates for the effects tended to be more positive for smaller scholarships and closer to zero for larger scholarships, despite the lack of statistical significance in some cases. On one hand, this can be taken as evidence that the monetary value of a scholarship award has a negative association with the enrollment rate. A more plausible explanation, however, is that that students who are of higher merit (in this case, those who receive larger awards) are less swayed by scholarships when making their college decisions.

Table 3.4: Estimates of effects of scholarship receipt on student enrollment when including interaction terms.

Disc.	Bandwidth	Estimate	95% CI	p-value
$S0 - S1$	By Count	0.076	(-0.079, 0.230)	0.33
	By Ability	0.228	(-0.037, 0.493)	0.09
$S1 - S2$	By Count	0.141	(0.019, 0.263)	0.02*
	By Ability	0.099	(-0.055, 0.252)	0.21
$S2 - S3$	By Count	0.009	(-0.173, 0.192)	0.92
	By Ability	-0.053	(-0.322, 0.216)	0.69

* $p < 0.05$

Existing literature on this topic has seen mixed evidence with some authors arguing that price responsiveness declines with ability (e.g. [50]) with others arguing that academically gifted students respond strongly to scholarship receipt (e.g. [20, 116]). In addition, students from higher-income families tend to be less sensitive to financial aid offers and college tuition costs [103, 111]. Thus, any association between academic achievement and family income must also be taken into account when examining these results [65].

The calculated 95% confidence intervals are also quite wide, particularly when estimating effects based on bandwidths optimized based on ability. As seen in Table 3.2, these ability-based bandwidths contained far fewer students than their count-based counterparts. As such, the corresponding confidence intervals were quite broad. That said, despite its broad confidence interval, the effect estimate across the $S0-S1$ discontinuity had the second-lowest p-value (0.09) among all estimates.

We also calculated the effect estimates when including all available covariates. In general, adding covariates did not significantly alter any estimates, regardless of whether looking at the inclusion of interaction terms or whether the bandwidths were optimized by count or ability. This speaks to the validity of the RDD design and the assumption of local

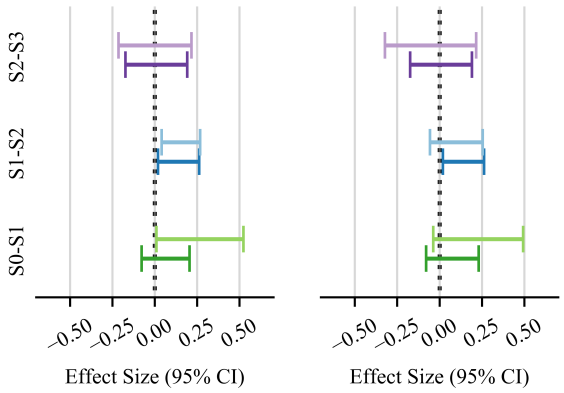

Figure 3.4: Estimates of the effect of scholarship on the enrollment of students at the University. The top, lighter colored bars represent effect estimates based on bandwidths optimized in terms of ability. The bottom, darker colored bars represent effect estimates based on bandwidths optimized in terms of counts of students. Estimates without interaction terms are on the left and with interaction terms are on the right.

randomization. If other covariates were in fact associated with either the assignment or outcome variables, the resulting effect estimates would vary wildly from the estimates when excluding covariates. In this case, any differences we saw in both the effect estimates and the associated confidence intervals were relatively inconsequential.

3.5.3 Limitations

This work considers the receipt of any scholarship award in a binary manner. Approaching scholarship awards in this manner disregards the difference between the awards themselves. For example, The difference in monetary value between $S0$ and $S1$ is identical to that between $S1$ and $S2$. However, the difference in monetary value between $S2$ and $S3$ is significantly larger, as shown in Table 3.1. Treating each of this continuities in a binary manner ignores such differences in their monetary value. However, it must be noted that there were no differences in award amounts at each of the discontinuities - there was a single award amount for those immediately to the left of the discontinuity and a single award amount for those immediately to the right of the discontinuity, with "left" and "right" referring to students'

rankings based on academic profile. Without any variability in award amount, it becomes extremely difficult to disentangle the effects of the receipt of an award across a discontinuity from the award amount. In allocating the scholarships, we were not at liberty to have varying award amounts for the same academic profiles.

One potential way around this is to use a model specification that takes into account both scholarship award amount and scholarship receipt. This could be of the following form:

$$Y = \alpha + X_1\beta_1 + S1\gamma_1 + X_2\beta_2 + S2\gamma_2 + X_3\beta_3 + S3\gamma_3 + \epsilon \tag{3.6}$$

where $S1$, $S2$, and $S3$ refer to the respective award amounts and X_i refers to the receipt of award i. Of note is that this functional form does not examine each discontinuity in isolation but instead looks at them concurrently. Using such a functional form without additional interaction terms assumes that there is no variability with respect to other covariates across ability. Additionally, because it does not use a regression discontinuity design, the local randomization assumption does not hold and stronger assumptions must be made regarding the data. As it is, we did not feel comfortable making such assumptions for this analysis but do intend to explore this idea further as we investigate the effect of other covariates on enrollment.

Another limitation of this work is that it is restricted to the enrollment decisions at a single institution for a single year. As Goodman notes, evaluating the effect of merit-based aid is largely dependent upon understanding how students, particularly higher ability students, respond to changes in the costs of their college options [73]. The effect of scholarship receipt and an increase in financial aid award will vary based not only individual characteristics but also characteristics of the awarding institution as well as those of other institutions being considered by the student [184]. However, we do not have any information on alternative college options available to students, thereby making it difficult to distinguish the effect of the University's offer compared to those of other schools. Some previous studies (e.g. [64, 152]) have specifically collected information on which alternative institutions students

attended. We did not have access to such data but having it could present a more complete picture of the student enrollment decision.

3.5.4 Conclusions

In this work, we used a regression discontinuity design to estimate the effect of a merit-based scholarship awarded to out-of-state students as a tuition waiver by a large, public US university in order to entice students to enroll. The disbursement of the award we examined was previously optimized by our group to maximize the enrollment of students. Thus, we had complete visibility into the allocation/selection process as well as assurance that students were not able to manipulate the assignment process in their favor. We examined three distinct discontinuities across four award amounts for the same scholarship award. We used bandwidths for our study design that are optimized in terms of student ability and in terms of the number of students. We find some weak evidence that the scholarship awards increased enrollment at the awarding institution. In general, trends among enrollees appear to indicate that students with higher academic achievement are less swayed in their enrollment decision by tuition waivers compared to students of lower academic achievement.

As we noted in our work discussing our numeric optimization process [11], the disbursement strategy for the scholarship examined had previously been developed by an external consulting service. Due to their proprietary techniques, the University had limited visibility into the approaches used in determining award allocations. This reduced visibility ultimately limited the degree to which the University could examine the causal effects of their scholarship policies with respect to their intended outcomes. Bringing the optimization process under the purview of the University allowed us to have visibility into the scholarship allocations and employ a study design that leveraged this fact.

We are now working with University administrators on how to best use/interpret our results for future iterations of the scholarship. Ultimately, we would like to develop a scholarship allocation strategy that maximizes the impact the scholarships have on student enrollment decisions across the academic achievement spectrum.

3.6 Publications/Presentations

- None thus far (publication in draft)

Chapter 4

PREDICTING FIRST-YEAR UNDERGRADUATE ATTRITION

4.1 PREAMBLE

This Chapter is the first of two to examine students after their post-secondary enrollment. The ideas for this Chapter were actually conceptualized during my first year as a PhD student, inspired in part by churn prediction in telecommunication networks. Since then, the project has gone through many iterations over the years. In the earliest form of this work, we tried to examine attrition after only a student's first term on campus. In its latest form, we focused on examining whether student attrition can be predicted using only students' first year of information. I would hope that its final form will be as part of a larger system to help alert advisors and administrators of students at risk of drop out. What's more, in this work, we tried to give more practical guidance for those working in institutional settings by examining what types of data are most useful for understanding attrition. From a professional perspective, this work was also what first drew interest from University of Washington administrators and staff in my work.

4.2 Abstract

Each year, roughly 30% of first-year students at US baccalaureate institutions do not return for their second year and billions of dollars are spent educating these students. Yet, little quantitative research has analyzed the causes and possible remedies for student attrition. What's more, most of the previous attempts to model attrition at traditional campuses using machine learning have focused on small, homogeneous groups of students. In this work, we model student attrition using a dataset that is composed almost exclusively of information routinely collected for record-keeping at a large, public US university. By examining the

entirety of the university's student body and not a subset thereof, we use one of the largest known datasets for examining attrition at a public US university (N = 66,060). Our results show that students' second year re-enrollment and eventual graduation can be accurately predicted based on a single year of data (AUROCs = 0.887 and 0.811, respectively). We find that demographic data (such as race, gender, etc.) and pre-admission data (such as high school academics, entrance exam scores, etc.) - upon which most admissions processes are predicated - are not nearly as useful as early college performance/transcript data for these predictions. These results highlight the potential for data mining to impact student retention and success at traditional campuses.

4.3 Introduction

Student attrition has long been a topic of great interest in higher education research, with government reports on attrition dating back over 100 years [169]. This interest stems from the fact that students who do not graduate are a lost investment on many fronts. For higher education institutions, limiting attrition is central to their financial sustainability as they devote scarce resources towards classes and services for non-completing students [91]. In particular, it is estimated that 30% of United States (US) first-year students do not return for their second year of post-secondary education with US taxpayers spending nearly $2 billion annually on educating non-returning first-year students alone [148]. Institutions are also concerned with attrition rates because they are central to estimates of institutional effectiveness, thereby affecting funding opportunities and government support [84]. Highlighting the impact of attrition at the institutional level also says nothing of its impact on students, who devote time, effort, and finances towards unfinished educational pursuits. Leaving college drastically alters career trajectories for students and those without college degrees face continually declining job growth and worsening job prospects [39].

In light of this, understanding motivations for students to drop out and possible remedies thereof is of great importance [57]. Empirical evidence to build student attrition theory has traditionally focused on survey-based research [163, 35]. However, survey instruments are of-

ten costly to implement, time-consuming for data collection, and produce results that are not always generalizable across institutions due to vastly different student profiles [175, 33, 35]. Institutional data that is routinely collected at colleges and universities (e.g. student application and transcript data) can provide an alternative data source and a way to supplement survey-based measures [35]. Leveraging data sources already in existence can add a means to more efficiently examine the student attrition problem and help institutions remedy the issue of attrition. One field that is primed to take advantage of this institutional data is educational data mining (EDM) and its focus on data-intensive techniques in educational settings [138, 18].

EDM is an emerging field with much of its research on attrition centered on massive online open courses (MOOCs) and other online environments (e.g. [190, 77]). Studying attrition in MOOCs and other online settings lends itself to expansive data collection opportunities and a detailed monitoring of students [125]. This limits the extent to which this work can be generalized to more traditional campus settings (i.e. campuses where learning is primarily on-campus, in-classroom). Meanwhile, EDM-centric work on predicting attrition at traditional campuses has been scarce and usually limited to small, homogeneous subsets of students rather than the entirety of a college student population. Additionally, the focus when predicting attrition is usually on how well it can be predicted and less so on what type of data is best for these predictions.

In this work, we predict the attrition of a large number of undergraduate students (N = 66,060) using only their first year of academic data. The students we examine are not from a single department or major within a university. Rather, they span the entirety of a student body, thereby comprising a dataset with heterogeneous aspirations, backgrounds, and goals. In addition, we rely almost entirely on data that is routinely collected at institutions of higher education. With this data, we seek to answer two questions: to what extent can undergraduate student attrition be predicted using a limited amount of data from registrar records and what types of data from registrar records are most useful in predicting attrition. The first of these has been explored in the past while using smaller and/or homogeneous

student populations; the second has not been systematically examined in the literature to our knowledge.

To answer the above questions, we mine the institutional data records at a large, public university in the US and engineer features for predictions. We then create numerous machine learning models using the engineered features and compare the performance of these models to each other. Then, we create separate machine learning models using only groups of features and not the entirety of the feature space to compare the predictive power of different subsets of institutional data. This work is an extension of our previous work on modeling student attrition using a limited amount of data [14] but where we previously focused on using the *first term's* data in generating features for prediction, we use the *first year's* in this work. We also extend our previous work to build additional machine learning models, predict attrition as defined according to two different definitions (overall graduation and re-enrollment after students' first year), and examine the types of feature subsets most useful in predictions. In so doing, we present two key findings, both of which have many implications for administrative policy in higher education:

- We demonstrate that the graduation and second-year re-enrollment of students can be predicted using data that is routinely gathered at institutions of higher education.
- We show that demographic and pre-entry features have less predictive power than data on student academics.

4.4 Related work

There are many examples of predicting attrition at traditional campuses. Most of these focus on small, homogeneous subsets of students. Moseley predicted the graduation of 528 nursing students using rule induction methods, obtaining high accuracies but not controlling for the number of terms/semesters examined for each student [121]. Dekker et al looked at only the first semester grades of 648 students in the Electrical Engineering department at the Eindhoven University of Technology and were able to predict dropout with 75-80% accuracy [55]. Kovačić used tree-based methods on a similarly-sized dataset of 453 students at the

Open Polytechnic of New Zealand, finding ethnicity and students' course taking patterns to be highly useful in prediction [99]. Bayer et al. looked at 775 applied informatics students at the Czech Republic's Masaryk University across three years [24]. Without limiting the amount of information available for each student, they found that including features related to students' social behavior can boost prediction accuracy by over 10% for some models. These and similar studies, however, focus on relatively small (e.g. $N < 2{,}000$) subgroups of students with similar academic pursuits/foci. In addition, there is little consistency with respect to the timeframes across which data is examined for each student. Other approaches to predict attrition at traditional campuses include early alert systems, which are often labor intensive and poorly funded [157]. These alert systems have been shown to positively benefit students (e.g. [90]), but usually rely on data gathered in the midst of a course or an academic term (e.g. [144, 87]), which may not always be feasable.

The work we present more closely relates to a subset of literature looking at student attrition in the context of the heterogeneity of students across an entire campus and not just a subset thereof. Our work also deals with much larger student populations than those described above and, in this sense, it more closely resembles a more recent body of literature. Delen used 8 years of institutional data on over 25,000 students at a large, public US university, predicting whether the students would return for their second year [56]. However, due to class imbalances, Delen re-sampled the majority class and ultimately used only 6,454 students for predictions. Ram et al. used data on about 6,500 freshmen at a large, public US university to predict whether students would drop out after their first semester, and for those that did not, whether they will drop out after an additional term [136]. Ram et al. supplemented data from institutional databases with student smart card transactions to infer social integration. More recently, Nagy and Molontay predicted the dropout of 15,825 students from the Budapest University of Technology and Economics using only their information prior to college entry with some success [123].

There are a few ways in which our work contributes to this body of literature. Firstly, we use a much larger dataset than has been previously examined specifically for attrition

(66,060 students). We examine the entirety of a large university's student body and we do not limit the extent of heterogeneity of the students in the dataset. Additionally, we also address the question of what types of features are most useful in predicting student attrition. In particular, previous works have generally used all available data sources concurrently in determining which students will attrite. In this work, we explore what types of routinely-collected institutional data fare best when predicting attrition by comparing performance using different data subsets in isolation. Finally, we concurrently compare predictions for two different definitions of "attrition," highlighting the degree to which operationalizing the term can impact results.

4.5 Methods

We describe the methods for this work by first detailing the data used in the project. We then give relevant operational definitions with respect to how we define attrition. Thereafter, we discuss the data subsets used in the predictions and the features generated. Lastly, we describe the setup of the machine learning experiments.

4.5.1 Data Description

We collected psuedonomyized, de-identified data from the University of Washington (the University) data stewards in 2017. The University is a traditional campus setting where a vast majority of instruction is in person and face-to-face. No personally identifiable information was collected for the students; instead, students were referenced using unique identifying keys. Table 4.1 shows the tables that were pulled from the registrar databases. In general, the data included information on students' demographics, complete transcript records at the University, and information from applications to the University. We did not have any information on students' financial aid status or economic status other than that which was derived from their ZIP code, as described below. Socioeconomic factors can play a large role in the student attrition process [32], however, we did not have access to student finances for use in this work. We also did not have access to any exit surveys from students who had

either left the University or had graduated.

Table 4.1: Data pulled from registrar databases

Table	Description
Application Data	Information from student applications to the University including high school coursework
Guardian Data	Information on student guardians as pulled from student applications to the University
Demographic Data	Information on student demographics including date of birth, race, ethnicity, gender, etc.
Major Data	Information on majors declared by students on a term-by-term (quarter-by-quarter) basis
Test Score Data	Information on student standardized test results
Transcript Data	Information on student coursework and grades on a term-by-term (quarter-by-quarter) basis

We restricted data to high school graduates who first enrolled at the University as matriculated, baccalaureate-degree-seeking undergraduate students between 1998 and 2010 without previously attending another post-secondary institution full-time. These students are henceforth referred to as "freshmen." The dataset included students who were in a college in high school program but excluded those who attended junior/community college full-time after high school and then transferred to the University. Because the data was pulled in 2017, we used the year 2010 as a cutoff to allow for six full years of visibility on student academics at the University before labelling a student as a "non-completion," as defined in Section 4.5.2. In total, the dataset consisted of 66,060 unique freshmen entrants. We then further limited the data for each student to information through one calendar year from each student's first enrollment at the University. This data was limited to one calendar year for all students, regardless of the number of courses they took/passed, their grades, or their backgrounds.

After joining tables of interest using the unique student identifiers, we created features for the prediction experiments by either pulling them directly from the raw data or engineering

them for each student. The features were grouped in 7 groupings, which are described in Section 4.5.3; a comprehensive list of features and descriptions thereof is available upon request but was not provided in this writing in the interest of space. In total, there were 1,405 features and all features were generated for each student without exception.

4.5.2 Definitions

Ambiguity with respect to operational definitions of dropout in literature on student attrition can make it difficult to compare results across studies [127, 174]. There are numerous ways in which attrition has been defined in existing literature, be it students dropping out from a particular course (e.g. [121]), re-enrolling after their first term (e.g. [3]), re-enrolling after their first year (e.g. [56]), graduating on time (e.g. [14]), or reaching some other relevant milestone (e.g. [55]). In this work, we defined attrition in two ways and analyze both. We examined attrition from students' first year to their second ("re-enrollment" and "non-re-enrollment") as well as looking at whether a student graduated on time ("graduate" and "non-completion"). We do not examine attrition on a term-by-term basis because of the relatively few students who leave the University after only a single term, as discussed in Section 4.6.1. We operationally defined non-completion and re-enrollment as described below.

Non-Completion

We defined "non-completion" as any freshman student who did not graduate with a baccalaureate degree from the University within 6 calendar years of first entry to the University. We defined a "graduate" as a freshman who graduated from the University with a baccalaureate degree within 6 calendar years of first enrollment. The University uses a quarter term system and we used the span of four consecutive academic quarters as a measure of one calendar year. Six calendar years for graduation was thus the span of 24 consecutive academic quarters. This definition of non-completion only accounted for students' first baccalaureate degree and did not take into account double-majors or double degrees. For example, if a

student was simultaneously pursuing two baccalaureate degrees but only graduated with one in five years, they would be a graduate; alternatively, if the student had graduated with both degrees but during their seventh year, they would be considered a non-completion. Because we focused on registrar records from a single institution, defining non-completion in this manner does not take into account students' academic progression after leaving the University. This is because we only had access to registrar records from a single institution and did not track students across multiple institutions - they could have very well transferred from the University and graduated in good standing.

We accounted for students who took part in a college in high school program by converting their transferred credit total to a count of academic quarters completed while assuming typical full-time enrollment at the University. For example, if a student completed 30 credits in a college in high school program, we converted this credit total to a count of terms completed at the University (in this case, 2, as students typically take 15 credits per term). We rounded the result from this conversion where appropriate. We then deducted this number when determining whether the student had graduated within an appropriate amount of time.

Re-Enrollment

We defined "re-enrollment" as a student who completed at least one additional course within one calendar year of the end of their first calendar year at the University (i.e. within 4 academic quarters from the end of their first year). "Non-re-enrollments" were students who were not re-enrollments. In this work, the definitions of graduation and re-enrollment were treated mutually exclusive in that all graduates were not necessarily re-enrollments. It should be noted that the University requires students who do not enroll for two consecutive terms without an excused leave to be re-admitted at the discretion of the University.

4.5.3 Feature Groupings

For every student, we engineered the subsets of features that are described below. For all student grades, we calculated a grade percentile and a z-score by comparing each students' grades to the grades of all undergraduate students who had taken the same course at the same time. References to grades include the student's GPA (on a 4.0 scale), their percentile score (from 0-100), and their z-score for courses (representing the number of standard deviations from the mean, assuming a normal grade distribution). References to "performance" for the feature groupings include grades and credits earned, at the least. In some cases, references to performance may also include the number of graded credits earned (versus courses taken pass-fail) and the number of credits attempted. A brief description of each of the feature subsets is provided in Table 4.2.

Table 4.2: Data subsets used in predictions

Subset	Description
Base Data	Year and quarter of University entry (included with every other data subset)
Demographic Data	Non-academic data prior to entry to the University, including demographics
Department-level Data	Measures of performance aggregated by course department
First-Year Summary Data	Aggregated measures of academic performance during first year
Grouped Course Data	Measures of performance aggregated by course number and STEM gatekeepers
Major Data	Counts of majors declared on a term-by-term basis
Pre-Entry Data	Academic data prior to entry to the University.

Base Data

Base data consisted of only three features and was included in the feature space when making predictions using every other data subset described. The base data included students' calendar year of entry to the University, their quarter of entry to the university (i.e. which of the four academic quarters was a student's first; ranging from 1 to 4, with 1, 2, 3, and 4 corresponding to winter, spring, summer, and autumn academic quarters, respectively), and a quarter-year variable which consisted of students' year of entry multiplied by 4 and added to the quarter of entry to create a relative time scale. These features were included to account for any time-related variation in graduation rates.

Demographic Data

Demographic data consisted of student's non-academic information prior to entry to the University. This included, but was not limited to, students' gender, race, ethnicity, age at college enrollment, veteran status, and student athlete status. We also included information from students' application to the University, such as information on the students' high schools (excluding high school grades), parents' educational attainment, and students' ZIP (postal) code, which was either pulled from their high school information or, when unavailable, from their university application. We joined students' ZIP codes with 2015 US census data[1] to find the average income and educational attainment in each ZIP code. We also included the distance from the University to each student's home ZIP code. Features derived from ZIP codes were the only features from sources external to the University's registrar databases.

Department-level Data

Department-level data consisted of student performance in course offerings grouped by course prefix. For example, this included performance in all BIOL (biology) courses grouped together, performance in all HIST (history) courses grouped together, etc. We excluded course

[1]From the US Census Bureau's American Fact Finder

prefixes wherein at least 10 students from the dataset did not take a course. In all, this included 200 unique course prefixes and 1000 features, with GPA, percentile grade, z-score, credits earned, and graded credits earned calculated for each prefix. We used department-level data instead of individual course data after preliminary modeling using individual courses did not yield strong results. The expansive feature space when engineering features across individual courses also significantly increased the requisite computational power/time for modeling and we decided against pursuing this further.

First-Year Summary Data

First-year summary data consisted of aggregate measures of students' first year at the University. This included, among other things, students' course performance, credits taken, number of courses failed, number of quarters enrolled, and enrollment in a freshman seminar courses. The first-year summary data also included aggregate measures of students' performance in their first, second, third, and fourth quarters as well as student performance in the last academic quarter for which they were enrolled during their first year (regardless of which quarter it was). We also included differences between students' performance in successive quarters.

Grouped Course Data

Grouped course data consisted of student course performance grouped either by course number or by performance in "STEM gatekeepers." To group courses by course number, we aggregated performance across all courses that were numbered below 100, from 100-199, from 200-299, from 300-399, and 400+. The course numbering generally reflected whether the course was designed to be taken by lowerclassmen or upperclassmen and, in some cases, also indicated during which year students typically took the course. STEM gatekeepers refer to introductory science, technology, engineering, and math (STEM) courses which often function as pre-requisites for STEM majors and degrees. These gatekeeper courses tend to be highly competitive and performance in these courses is a key determinant of whether a

student will be accepted into any of the highly competitive STEM majors. We grouped the performance in STEM gatekeepers by course department and topic (e.g. the calculus series, the general chemistry series, the organic chemistry series, etc.) as well as across all STEM gatekeepers.

Major Data

Major data consisted of counts of students' major declarations during their first academic year. In most cases, students entered the University with a "pre-major" designation before declaring their major(s) of interest some time during their first or second year. These pre-major designations varied based on field of interest (e.g. pre-engineering, pre-nursing, pre-health, etc.). Students' majors were recorded on a per-quarter basis by the University (once per quarterly transcript record) and we tallied the counts of major declarations for each student across the entirety of their first year. For example, a student who declared a math major in their first two quarters only to switch to geography in their third quarter and then add a history double major in their fourth quarter would have the values 2, 2, 1 in the math major, geography major, and history major features, respectively.

Pre-Entry Data

Pre-entry data consisted of students' academic information prior to attending the University. This included, among other things, students' entrance exam scores, high school GPA, high school coursework, and college in high school program participation and performance. We did not include any information on students after their enrollment at the University in the pre-entry data.

4.5.4 Machine Learning and Predictions

We randomly divided the students into training and test sets using a 80-20 split (N in training = 52,848; N in test = 13,212). We used the same test set when evaluating the predictive

performance of each of the models to allow for direct comparisons to be made. The data was highly skewed with graduates and re-enrollments comprising 78.5% and 93.1% of all the data, respectively. Graduates and re-enrollments comprised 78.0% and 92.9% of the test data, respectively. Though dealing with class imbalances is of great interest when examining freshmen attrition [172], we did not use any balancing techniques as we wanted to work with the data in its original, unaltered form. We scaled the training data by subtracting the median of each feature and dividing by the respective feature's interquartile range. We subsequently scaled the test data using the scaling values for each feature from the training data.

We used five different machine learning models to predict each student's graduation and re-enrollment: regularized logistic regression (LR), K-Nearest Neighbors (KNN), random forests (RF), support vector machines (SVM), and gradient boosted trees (XGB). We trained each model across the entirety of the training data and used the same training instances to train each of the models. We trained each model separately to predict graduation and re-enrollment. We tuned model hyperparameters for each model using 5-fold cross validation on the training data, after which the models were re-trained on the entirety of the training data using the tuned hyperparameters. We report final error metrics and performance on the test set, which was consistent across all models, regardless of whether predicting graduation or re-enrollment.

After developing predictive models using all features, we created regularized logistic regression models using each of the 6 feature subsets highlighted in Section 4.5.3 in isolation. The base data (see Table 4.2) was included in the feature space for each data subset. The rationale behind using regularized logistic regression for these models is further discussed in Section 4.6.3. We understand that an alternative approach would be to test all the models listed above for each of the data subsets to find the best performing model/subset combinations. That said, we believe our approach was still suitable for comparing different data subsets. When modeling using data subsets, we used the same observations as before to train each of the models and, as before, we developed a separate model for predicting graduation

and re-enrollment for each of the data subsets. As such, the *training instances* were the same across models but the *training features* differed depending on the feature subset used. We tuned the regularization strength for these regularized logistic regression models using 5-fold cross validation on the training dataset and we report results on the test set.

4.6 Results and Discussion

4.6.1 Student Characteristics

We show the number and proportion of graduates and re-enrollments in Figure 4.1. In all, 78.5% of students were labelled graduates while 93.1% of students were labelled re-enrollments. These proportions were verified with the University's office of institutional analysis. Such highly skewed data towards graduates and re-enrollments can be expected in a large, tier-1 research university setting where there has been considerable, long-standing effort to improve the overall attrition rate over time. That said, it must also be noted that at an institution with such a large student population, even small fractions of the student body represent hundreds of students on an annual basis. Across the timeline of the dataset (13 cohorts), 14,196 non-completions and 4,593 non-re-enrollments represent 1,092 and 351 students on an annual basis, respectively.

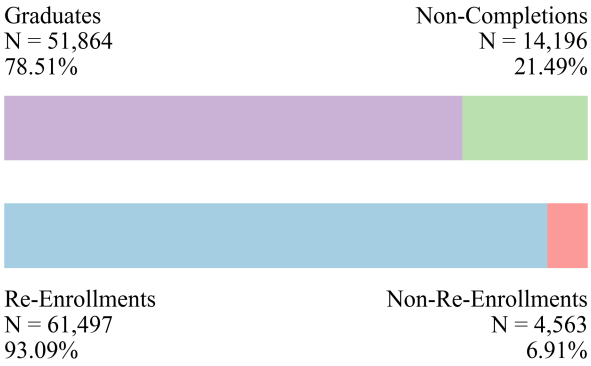

Figure 4.1: Counts and percentages of classes in the dataset. Definitions are provided in Section 4.5.2.

We show the cumulative percentage of students who either graduated or left the University across time in Figure 4.2. We used the first year as a cutoff for the data because, historically, a large number of students decide whether they will continue with their higher education pursuits during and immediately after their first year [148]. As such, developing models that can predict whether students will re-enroll for a second year and whether they are on a trajectory towards successful graduation could help administrators and academic advisors more effectively develop and deliver interventions directed towards students in need of assistance. When examining the data, 27.5% of all non-completions leave the university prior to the start of their 2nd year, 51.9% of non-completions leave the University between their 2nd and 6th year, and 20.6% continued to be enrolled at the University after their 6th year. The difference in number between non-completions who did not return for their 2nd year and non-re-enrollments can be attributed to non-re-enrollments who later returned to the University and graduated on time. Less than 5% of non-completions and less than 15% of non-re-enrollments left the University after only one term, leading us to not examine attrition after the first and second terms. In settings where attrition rates are higher after students' first and second terms, it may be more relevant to examine the performance of classifiers after one or two terms.

Figure 4.2 also shows that a majority of graduates (65.6%) completed their degrees during their fourth year at the University. The mean and median completion time for all graduates was 16.6 and 15.0 calendar quarters, respectively, from first enrollment. This is particularly apparent due to the near-sigmoidal shape of the cumulative graph for graduates, with a sharp rise during students' fourth year. We also see that there is a relative lack of students who graduated prior to the start of their third year. This highlights the difficulty in predicting graduation based on students' first year - a student typically does not graduate until several years later, during which a host of influences can shape an academic trajectory, be they personal, financial, or academic.

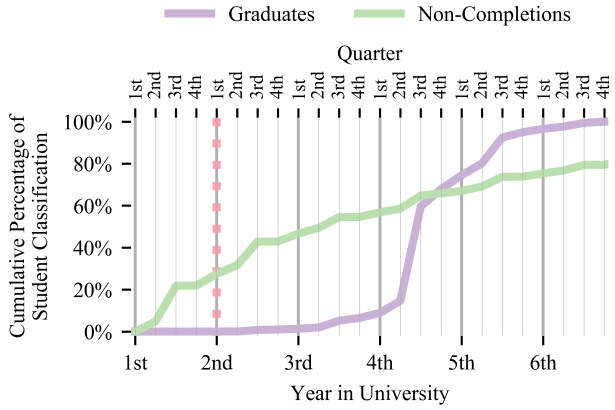

Figure 4.2: Cumulative graduation and non-completion curves of students. Years and quarters are relative to the time of first enrollment. The dotted line indicates the point to which data is limited for each student. Only students' first six years are shown, per the definition of "graduate."

4.6.2 Predictions Using Different Algorithms

Table 4.3: Prediction results using all data features. Baseline values are based on test set.

	Graduation		Re-Enrollment	
Model	Accuracy	AUROC	Accuracy	AUROC
Baseline	78.0%	0.500	92.9%	0.500
LR	83.2%	0.811	95.0%	0.882
RF	83.1%	0.806	95.3%	0.887
XGB	83.0%	0.806	95.1%	0.885
KNN	82.5%	0.798	94.8%	0.876
SVM	78.0%	0.780	92.9%	0.862

We show the performance of each of the models using the entirety of the feature space in Table 4.3. The baseline measure in the Table refers to the majority class compositions in the test set. Generally speaking, most of the models had a similar comparative performance for each prediction task (i.e. predicting either graduation or re-enrollment). This hints at an effective ceiling with respect to predictive power from the types of features being

used (i.e. ones pulled from registrar records) and that additional representations of the student experience (be they academic or social) should be incorporated. Alternatively, a more complex predictive model (e.g. deep neural networks) may also fare better in making these predictions. That said, given the data used, the models are able to predict the eventual graduation and re-enrollment of students fairly successfully, as evidenced by the relative improvements over baseline values for both prediction tasks.

For predicting graduation, logistic regression was the best-performing model, followed by random forests. When predicting re-enrollment, random forests performed the best, followed by gradient boosted trees and logistic regression. These results are generally in line with our previous work on similar tasks, where we found that logistic regression tends to work well compared to other models for predicting graduation and STEM attrition [7]. When examining the worst-performing models, the SVM model made predictions that consisted entirely of the majority class when predicting both graduation and re-enrollment, as seen by the models' accuracy being the same as the baseline values. Such results are typical of classifiers without much predictive strength on a dataset consisting of highly disproportionate classes. In this specific case, it may be remedied by using alternate kernels for the model, which we did not explore in this work.

We show the ROC curves for the models in Figure 4.3. These curves further illustrate the lack of differentiation with respect to model performance. For the same prediction task, the resulting ROC curves across the models were nearly identical with little difference in curvature. The more notable difference was when comparing the ROC curves for predicting graduation with those for predicting re-enrollment, as the curves for predicting re-enrollment were more prominently convex compared to those for predicting graduation. These curvatures, along with the metrics shown in Table 4.3, demonstrate that predicting students' eventual graduation is a more difficult task than predicting students' re-enrollment. We expected this as the cutoff for the data used in the predictions (i.e. students' first year) was near the point at which a student is classified as a re-enrollment (after their second year) but was much earlier than when a student was classified as a non-completion (after their sixth

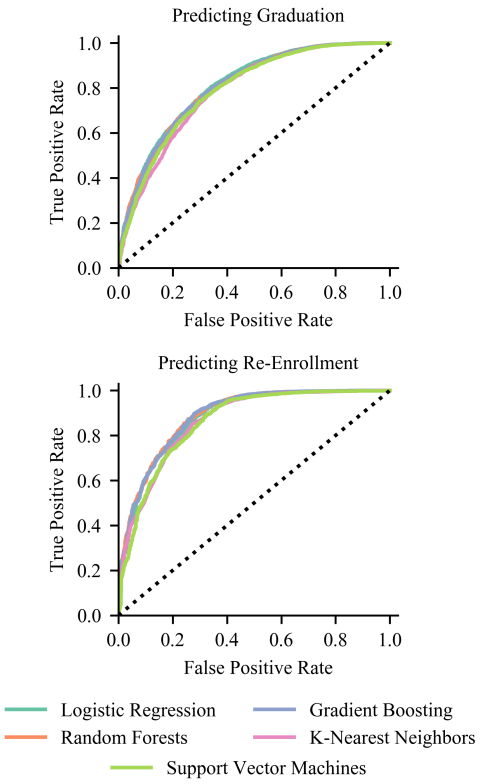

Figure 4.3: Receiver operating characteristic curves when using different machine learning models.

year). This helps highlight the degree to which differing operational definitions of attrition can vastly alter the perceived predictive strength of these classifiers. For other scenarios, alternate definitions of attrition may be more appropriate and the effectiveness of efforts to build predictive models will be colored by these definitions and institutional contexts.

We show the confusion matrices for the best models for predicting graduation and re-enrollment (logistic regression and random forests, respectively) in Figure 4.4. These matrices show a lower rate of false negatives for the models but a higher rate of false positives (i.e. students incorrectly classified by the models as having graduated or re-enrolled). To better understand this higher rate of false positives, we examined the complete transcript records of students who were classified accordingly. Across the false positives, we found numer-

ous instances of non-completions and non-re-enrollments who had left the University with relatively strong grades in comparison to their graduating and re-enrolling peers. These students also often appeared to be pursuing very competitive majors and/or appeared to have rigorous post-graduation plans (e.g. pre-medical and pre-dental students). Many of these students remained in a pre-major state prior to their departure, indicating that though they had relatively strong grades, they likely were not able to enter into their degree program(s) of choice for various reasons and had to leave the University to pursue these ambitions as a result. Unfortunately, the University does not have a centralized major application database for admissions and rejections to specific majors. Having so could shed light on much of the motivation behind these students' desire to leave the University and if it was, in fact, motivated by not getting into competitive majors. That said, the fact that many of these students were academically similar to their graduating and re-enrolling counterparts further illustrates why there appears to be an effective ceiling with respect to predictive power using the given data, as seen in Table 4.3.

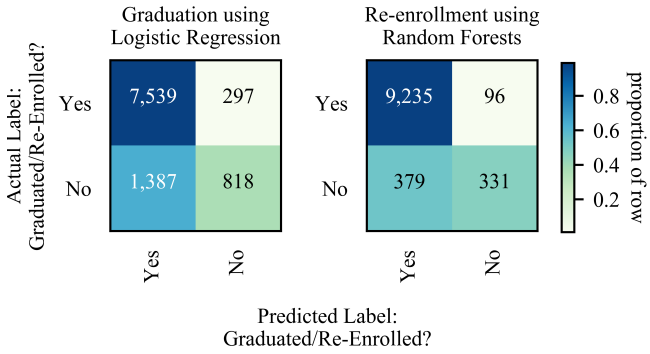

Figure 4.4: Confusion matrices when examining the top performing algorithms for predicting graduation (LR, left) and re-enrollment (RF, right).

From a practical perspective, it should be noted that the classification thresholds for these models were not tuned with respect to either sensitivity or specificity. In practice, when developing institutional systems to identify students at-risk of leaving, it may be useful to raise the classification threshold when predicting whether a student will graduate or re-

enroll, thus favoring lower recall at the expense of higher precision. This would effectively reduce the number of students who are predicted to graduate but in actuality do not (i.e. false positives) at the expense of more false negatives, which could be more acceptable when developing an alert system for students at risk of dropping out.

4.6.3 Predictions Using Different Data Subsets

After examining the results from predicting graduates and re-enrollments using all features, we used regularized logistic regression to predict graduation and re-enrollment using subsets of the data. We used logistic regression after we saw that it performed very well relative to other models for both prediction tasks (see Section 4.6.2) and because it had relatively fast training times due to having fewer hyperparameters to tune. This allowed us to more efficiently train the 12 different models that were needed when examining the performance of specific data subsets (i.e. separately modeling graduation and re-enrollment while using 6 different data subsets in isolation for each).

Table 4.4: Prediction results using specific data subsets. Baseline values are based on test set.

	Graduation		Re-Enrollment	
Subset	Accuracy	AUROC	Accuracy	AUROC
Baseline	78.0%	0.500	92.9%	0.500
All	83.2%	0.811	95.0%	0.882
FY-Sum.	83.0%	0.795	94.9%	0.855
Department	82.3%	0.788	94.6%	0.847
Grouped	82.5%	0.781	94.6%	0.845
Major	79.9%	0.661	94.2%	0.768
Demo	78.0%	0.634	92.9%	0.643
Pre-Entry	77.3%	0.630	92.9%	0.616

We show the results when using data subsets in Table 4.4 alongside the performance of the logistic regression classifier from Section 4.6.2. Transcript-based features tended to perform

better than information on students' prior to their enrollment at the University. More specifically, demographic data and pre-entry information did relatively poorly in predicting both graduation and re-enrollment. Intuitively, this is not a surprise as the admissions process at highly-competitive universities tends to be fairly selective with an emphasis on supporting and sustaining a successful yet diverse student body. Additionally, such institutions may already have efforts in place to reduce demographic disparities for student success. Meanwhile, when looking at transcript-based data subsets, first-year summary data performed the best with performance that was similar to using the entirety of the data. This is particularly noteworthy as the first-year summary data contained fewer features than the other transcript-based data subsets but was centered on summaries of performance across time rather than aggregations across course departments/numberings.

These findings are particularly interesting in light of work by other researchers. For instance, Nagy and Molontay found that attrition could be accurately predicted using what we outline as demographic and pre-entry features alone [123]. However, we do not see similar success here. We believe this could be due to vastly different educational settings and student profiles (e.g. here, most students tend to graduate/re-enroll while Nagy's student population primarily dropped out). In earlier work, Dekker et al. found that transcript-based features tend to have more predictive strength than pre-entry features, but examined this across rather limited data subsets [55]. Our results echo this finding. Recently, Manrique et al. found that attrition could be predicted using student performance in a few key courses [108]. Here, we find that aggregates across the first year tend to work better than more fine-grain representations of course-taking (e.g. grouping classes by course prefix and numbering). As discussed in Section 4.5.3, we decided against using individual course representations in this work.

We show the ROC curves for the regularized logistic regression models using each of the data subsets as well as the entire feature space in Figure 4.5. The fact that demographic and pre-entry data gave generally worse performance than transcript-based features is very much apparent from the ROC curves. Data on majors, meanwhile, tended to perform worse

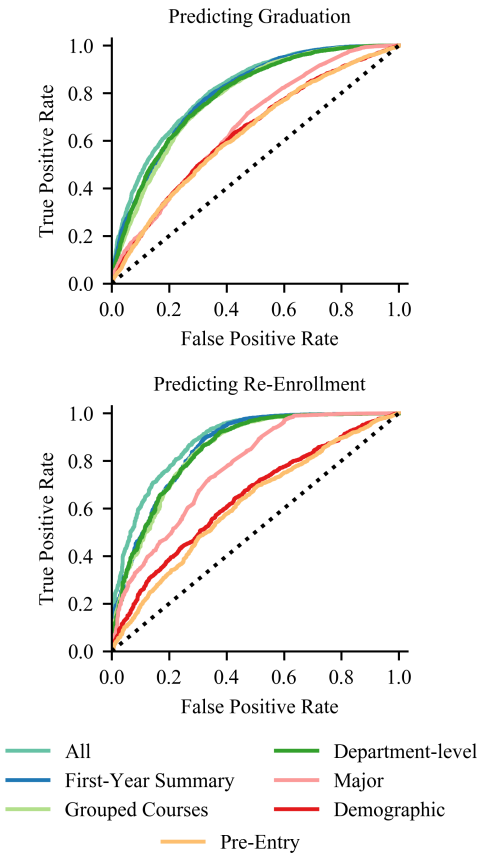

Figure 4.5: Receiver operating characteristic curves when using different subsets of data.

than other transcript-based features but better than demographic and pre-entry data. The fact that using data on majors did not yield particularly strong results likely relates to the fact that most students in the dataset were in a pre-major state across their first year and formally declared their major of interest later in their undergraduate careers. As noted above, a centralized major application system was not available, else it could have been leveraged in addition to data on majors to draw a more clear picture of student academic interest. The other transcript-based datasets, meanwhile, had very similar curvatures for the ROC curves when predicting both graduation and re-enrollment.

We show confusion matrices from using the best-performing data subset in Figure 4.6.

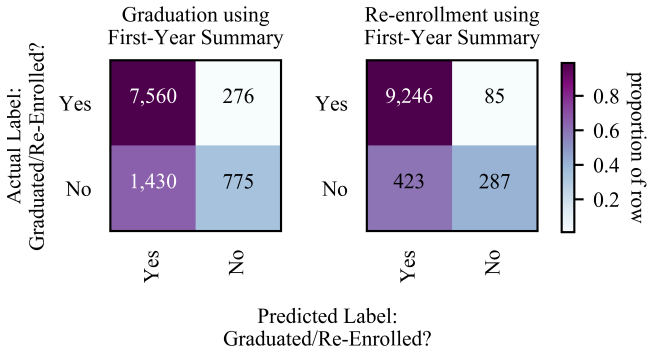

Figure 4.6: Confusion matrices when examining the top performing data subset for predicting graduation (left) and re-enrollment (right). The top performing data subset was the same for both tasks (first-year summary data).

The best-performing data subset for both prediction tasks was first-year summary data. By comparing these confusion matrices to those shown in Figure 4.4, it can be seen that using just a limited subset of features tends to classify the data similarly to models built on the entirety of the data. This is true not only in terms of how effective the models are in making predictions, but also with respect to the relatively high rate of false positives seen across all four matrices.

4.7 Future Directions

We believe the findings regarding the data subsets have wide-ranging policy implications, particularly for identifying students at risk of dropping out in large, public universities. In such settings, there may be longstanding effort to decrease demographic disparities with respect to attrition and, as a result, transcript records may be more viable as features in predictive models than pre-entry/demographic information. Furthermore, these settings may also be resource-constrained with respect to time available for staff to hand engineer features. In such settings, knowing which features would be most predictive of attrition without the need to hand-engineer features across the entirety of data available to institutions could save time and effort in building models. We have had conversations with administrators at

the University for better interpreting our results and improving the processes for identifying students in need of assistance.

Another direction of interest is better understanding the features used in predicting attrition. This includes not only further examining key individual determinants of attrition, as we have done in previous work [14, 7], but also finding the best combination of features across the subsets. We would like to examine this "minimum viable feature space" in the context of data available in registrar databases as well as investigate the degree to which these features relate to established theory on student attrition [57].

4.8 Conclusions

In this work, we use data from the registrar databases of a large, public US university to predict both graduation and re-enrollment using information limited to students' first calendar year at the university. We do this using a dataset of students that spans the entirety of the university student body and is thus much larger than previous studies predicting student attrition (N = 66,060). In so doing, we demonstrate that both graduation and re-enrollment can be effectively predicted using features generated from data that is routinely collected at institutions of higher education. Additionally, we also examine the degree to which specific subsets of registrar data can be useful in predicting attrition, finding that transcript-based features tend to outperform features based on student histories prior to college. This implies that effective strategies for intervention can be outlined based on registrar records.

Predicting re-enrollment after students' first year was a much more tractable task than predicting graduation. This can be attributed to the fact that predicting graduation necessitates predicting academic success years into the future from the point to which data was limited whereas predicting re-enrollment is within a much shorter timeframe. Considering the unpredictable influences that cause students to leave college prior to graduating (e.g. financial limitations, personal hardships, etc.), a more reliable prediction task may be to examine whether a student will return on a term-by-term basis. This could be particularly

useful to develop alert systems to identify students at risk of dropout. However, this was not explored in this work due to the relatively few students who left the University after a single term.

We found that there appears to be an upper limit for predictive power for our dataset. This demonstrates the limitations when relying solely on registrar data and shows the need for additional features on the student experience to improve predictive power. Some potential features of interest include measures of social integration on campus and of financial aid. Better understanding student aspirations beyond simply using declared majors could also be of interest, especially using alternate representations of student course-taking behavior, as shown recently by Luo and Pardos [106].

Lastly, we show that features generated from transcript records, particularly aggregates and summaries of students' academics, perform better for predictions than demographic and pre-entry data. Much of this is likely due to the selectivity of the University and its admissions policy. Nevertheless, it demonstrates how useful transcript data can be for such prediction tasks in contrast to information on students prior to college. We demonstrate that using subsets of data from registrar databases (in this case, aggregates of students' first year) can be nearly as effective for predictions as hand-generating a wide swath of features from different institutional data sources.

4.9 Publications/Presentations

- Lovenoor Aulck, Dev Nambi, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Mining University Registrar Records to Predict Undergraduate Attrition and Re-Enrollment. In *Educational Data Mining Conference*, 2019 ***(conference paper)***
- Lovenoor Aulck, Rohan Aras, Lysia Li, Coulter L'Heureux, Peter Lu, and Jevin West. Stem-ming the Tide: Predicting STEM Attrition Using Student Transcript Data. In *2017 KDD Workshop on ML4ED: Machine Learning in Education*, 2017 ***(workshop paper)***
- Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Predicting

Student Dropout in Higher Education. In *2016 ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*, 2016 ***(workshop paper)***

Chapter 5

UNDERSTANDING THE EFFECT OF FRESHMEN SEMINARS ON STUDENT PERFORMANCE

5.1 PREAMBLE

This final substantive research Chapter performs a program evaluation of a long-standing university program. The University of Washington's Office of First-year Programs oversees the FIG program and believed they were making a positive impact on students. There was some existing evidence for this by way of published studies that examined the University's FIG program specifically. However, the studies were conducted decades ago during the program's infancy and the 30-year old program had since grown to take a very different form. For this Chapter, I relied on the fact that FIG students were required to take a specific class to be part of a FIG, which allowed me to identify them from their transcript records. What's more, the FIG program had collected exit surveys from students but had never analyzed them - it was a rich dataset that had been collected without any real purpose in mind. I was provided this dataset and used it to further contextualize the impact of FIGs on student outcomes. In all, this chapter examines whether a program designed to positively impact student success does so by way of an analysis using both quantitative and qualitative approaches.

5.2 Abstract

Freshman seminars are a ubiquitous offering in U.S. higher education. Though these seminars have been evaluated in numerous studies, most studies have done so without employing matched comparison groups and using data at scale. In this work, we use data on nearly 58,000 students across 18 years at a public U.S. university to examine the impact of first-

year interest groups (FIGs) on student graduation and first-year retention. Using rich data from university databases and external sources, we apply propensity score matching to account for selection bias and confounding variables when comparing students. We find that graduation and re-enrollment rates for FIG students were higher than non-FIG students, an effect that was more pronounced for self-identified Hispanic students and self-identified under-represented minority students. Additionally, we analyze survey responses from over 12,500 FIG students to find that social aspects of the seminars, particularly making friends and knowing others taking the same classes, were the most beneficial to students. Interestingly, references to these social aspects were not disproportionately present in the responses of self-identified Hispanic students and self-identified under-represented minority students.

5.3 Introduction

Undergraduate retention has long been an area of great interest in education, motivated in part by consistently high rates of college students dropping out [57, 175]. Recent estimates from the National Center for Education Statistics (NCES) have about 40% of first-time, full-time bachelor's degree-seeking students at 4-year post-secondary institutions not graduating within 6 years of first enrollment [109, 93, 185]. The graduation rates are even lower for minority and Hispanic students [170]. These non-completing students account for a lost investment on many fronts, with students spending valuable time/energy on their unfinished educational pursuits and institutions collectively spending billions on educating the students who leave [91, 135]. Of particular interest is the fact that a large number of those leaving higher education without degrees are the 21-28% of first-year, full-time students seeking baccalaureate degrees who do not return for a second year of schooling [175, 128].

One way in which universities have combated freshman attrition and attempted to improve the college experience is through the implementation of freshman orientation seminars (freshman seminars). Freshman seminars are courses dedicated to helping incoming students transition to college life, both socially and academically [131, 183, 134]. The motivation behind freshman seminar courses relates to broader educational retention theory and the impact

of social integration on persistence [175]. Particularly on large campuses, freshmen often feel as though they don't have a personalized identity (i.e. as another "face in the crowd") while feeling overwhelmed by the competitive environment of higher education [168]. Freshman seminars (and, more specifically, learning communities) often rely on block scheduling and co-registration of classes. This promotes a sense of community, belonging, and provides freshmen with a means by which to more easily socialize and develop a peer group [177, 181]. This social engagement then helps students feel more connected to campus and more satisfied with the college experience overall [52] while making them feel less isolated as learners [177].

The popularity and ubiquity of freshman seminar courses has made them among the most studied course genre in American higher education [23, 126, 51]. However, the existence and effectiveness of these seminars on college campuses across the U.S. continues to be called into question [75]. Although some prior studies have used randomized controlled trials (e.g. [167]), large scale and causally rigorous studies of seminar effectiveness using matched comparison groups are rare [149]. Additionally, these seminars are still inadequately assessed, particularly with respect to practical considerations for their design [21] as well as the degree to which specific student demographic subgroups are differentially impacted by them.

In this work, we gather data from the institutional databases of a large, publicly-funded U.S. university (the University of Washington, UW) to examine the impact of freshmen seminars on student outcomes (namely, graduation rates and first-year retention). Using propensity score matching on nearly 58,000 students across 12 student cohorts (and 18 years of data), we use information on students prior to their post-secondary education to match students who enrolled in first-year interest groups ("FIGs," a type of freshman orientation seminar) with those who did not. We then examine the differences between these groups in terms of educational outcomes while also focusing on specific ethnic/racial groups (namely, Hispanic and under-represented minority students). To our knowledge, only one previous study examined the effects of any freshman seminar using a similar methodology but did so with only a few hundred students, finding freshman seminars positively impacted students' likelihood to re-enroll for a second year [42]. Here, we also examine graduation rates and

grades and do so across tens of thousands of students. In addition, we also examine specific student subgroups and the impact of FIGs on their academics.

To further understand the impact of FIGs on student success, we also examine the open-ended text responses of over 12,500 students who were asked what they found most valuable about their FIG experience. We first use unsupervised machine learning to develop a preliminary codebook and then manually tag each of the student survey responses. We then use these survey responses to better understand which aspects of the freshman seminar can be linked to student success, as we hope this will help inform future research, particularly with respect to peer groups in education. In addition, we believe this analysis of survey responses presents valuable insight into the effective design of these seminars.

In so doing, we present a large-scale analysis of a freshman seminar at a large, US university to determine its impacts on undergraduate success using a mix of quantitative and qualitative methods. This mixed methods approach relies on quantitative results from propensity score matching to evaluate the effects of FIGs on student outcomes and uses qualitative results from student surveys to better understand what students found most valuable about their FIG experience. We find that FIGs positively affect students in terms of graduation, re-enrollment, and grades. What's more, we find that students find social aspects of the FIGs, particularly making friends and knowing others in classes, to be the most beneficial to their success.

5.4 Related work

We present work to this study in the following manner: first, we provide a brief history of freshmen seminars. Then, we discuss prior studies that have assessed freshmen seminars. Lastly, we discuss prior works examining freshmen seminars specifically at the University of Washington.

5.4.1 History of freshman seminars.

Freshmen seminars take many different forms, including living-learning communities (LLCs), first-year interest groups (FIGs), and first-year experience (FYE) courses [134]. Though this work will focus specifically on FIGs, related research spans the spectrum of freshman seminar types. The freshmen seminar is defined by Barefoot as "a course intended to enhance the academic and/or social integration of first-year students by introducing them (a) to a variety of specific topics which vary by seminar type, (b) to essential skills for college success, and (c) to selected processes, the most common of which is the creation of a peer group" [22]. For the most part, these seminars tend to be smaller in size than most lower-division courses, thereby allowing for greater student-faculty interaction and an environment more conducive to developing and fostering peer relationships [126]. These relationships are believed to allow for greater social integration of students within a campus community, thereby increasing students' institutional commitment and their persistence towards the goal of graduation [178].

Freshman seminars have long been a part of the American higher education landscape, with orientation courses dating back to the 19th century [183, 62]. By the middle of the 20th century, a majority of institutions offered freshman seminar courses, with many campuses extending orientation beyond just a few days [62]. This trend, however, reversed through the 1960s due to concerns with universities giving credit to students as they adjust to college life [183]. A resurgence in seminar offerings was partly fueled by courses such as the University of South Carolina's "The University 101," which became a standard for freshman seminars as other campuses attempted to develop similar programs [149, 183]. By the early 1980s, a growing interest in freshmen seminars saw conferences on freshman orientation/seminar courses held at the University of South Carolina [183]. Since then, freshman seminars have become near-ubiquitous in the U.S. higher education landscape. The National Resource Center for The First-Year Experience and Students in Transition's most recent numbers report that freshman seminars are offered for credit at 87% of 4-year institutions, of which

52% require them for students [2].

5.4.2 Assessing freshman seminars.

Freshman seminars are widely believed to positively impact student retention, persistence, graduation, and academic performance [25, 126, 51, 128, 52, 66]. More specifically, numerous studies have found that freshman seminars tended to improve retention rates (recent examples include [149, 113, 131, 128, 166, 189]), improve graduation rates (recent examples include [98, 113, 128, 150]), and improve grades (recent examples include [131, 89, 166, 189]), among other outcomes.

Despite these outputs, however, it is difficult to disaggregate which aspects of freshman seminars contribute to specific student outcomes [133] and much of this is because of the study designs employed. Most studies examining freshman seminars rely on quasi-experimental frameworks, which do not explicitly account for selection bias [52]. In addition to quasi-experimental frameworks (e.g. [113, 98]), studies have also used meta-analytic approaches (e.g. [25, 131]) and/or regression approaches (OLS, multilevel, or other) (e.g. [126, 166, 189, 89, 133]) in their assessment of freshman seminars. These studies compare students without matched controls or comparison groups, which have been under-utilized when examining the effects of freshmen seminars [149]. Examples of more causally rigorous studies on freshman seminars have used randomized controlled trials (e.g. [167]) or propensity score matching (e.g. [42]). However, these studies have also performed their analysis with smaller student populations and/or less expansive covariates in propensity score matching than those used in this study.

5.4.3 FIGs at the University of Washington.

The FIG program at the UW first began in 1987 and was modeled after a similar program at the University of Oregon [178, 168]. FIGs are optional for students and are not required for any graduation requirements; the FIG seminar only counts towards general education elective credit counts. FIGs are widely advertised to all incoming freshmen during orientation and

advising sessions, which are required for all incoming freshmen before they register for classes. FIGs are open for registration during the standard university course registration period to all incoming freshmen who have completed the requisite orientation sessions. The UW Undergraduate Academic Affairs' Office of First Year Programs (FYP) currently administers the FIG program and recent surveys of incoming freshmen by FYP indicate that only 3.6% of freshmen do not know what a FIG is by the time they start their coursework.

FIGs are presented as a cluster of classes, with participants co-enrolled in all classes of the cluster during their first academic term on campus. In addition to being co-enrolled in classes, all students within the same FIG cluster are also required to take a seminar class together. This seminar class has been led by an upperclassman peer since the program's inception and is focused on a discussion of students' personal experiences rather than academics, with an aim of developing a sense of involvement, participation, and community [181]. In the first year of the FIG program, 83 students enrolled in 4 different FIG clusters; in the second year of the program, 200 students enrolled across 8 FIG clusters; and in the third year of the program, 400 students enrolled across 20 different clusters [168]. Today, the FIG program has over 150 different clusters and enrolls between 50-60% of the UW's annual incoming freshman population, totalling over 3,000 FIG-enrollees annually. In the past, FIGs were organized around a central academic theme, such as pre-law or pre-engineering. Now, the program no longer explicitly delineates FIGs for particular interests, instead focusing on providing a more rounded experience for freshman entrants.

The UW operates on a quarter system and students in a FIG select a variable number of credits to enroll as part of a FIG, ranging from 2 credits (just the seminar class) to 17 credits (the seminar class plus 3 additional classes in which FIG-mates are co-enrolled) for their first quarter. The UW charges block tuition for all students who register for between 12-18 credits, which is considered full-time undergraduate enrollment. There is no additional charge when students take the FIG seminar course and there are no restrictions specific to the FIG seminar as to how students pay for their tuition charges. Additionally, there is also no additional tuition charge when a full-time student enrolls in a FIG, as FIGs consist of

a maximum of 17 credits. The above history and current state of the FIG program were confirmed with FYP.

Previous studies examining FIGs at UW looked at the program at its infancy in the late 1980s to early 1990s [181, 178]. Tokuno and Campbell looked at two freshman cohorts at the UW from 1988 and 1989 (about 6,660 students total), analyzing the effects of taking a FIG on scholarship and retention [181]. Examining the 1989 cohort, they found that students who enrolled in FIGs had higher overall retention by 1.6 percentage points and also had higher course completion rates than their peers. However, the study only compared rates of graduation among the student groups and did not explicitly address confounding variables, not the least of which is the selection bias associated with students choosing whether or not to enroll in a FIG. In addition, the study was conducted during the program's early years and less than 700 students had enrolled in FIGs during the program's first three years combined, thereby limiting the general scope of the work.

Tinto and Goodsell conducted a longitudinal study in 1991, examining about 440 FIG students across 21 FIG clusters as well as about 1800 non-FIG students [178]. In using both quantitative and qualitative approaches, they found that being involved in a FIG "seems to positively influence one's sense of one's own involvement as well as the nature of one's peers and general institutional climate," which then serve to impact both academic performance and persistence. As with the previous study, however, they did not account for selection bias among the students, particularly with respect to demographics and FIG students generally having higher entrance exam scores. Additionally, their comparison non-FIGs students were selected from specific courses and not across the entirety of the broader student population.

In this work, we compare FIG and non-FIG students by matching using propensity scores based on a rich set of covariates detailing student characteristics and high school backgrounds prior to their entry into higher education. Propensity score matching was employed because students selected whether or not to enroll in a FIG on their own, rather than being randomly assigned. Thus, comparing students across this non-experimental design poses the risk of treatment effects being biased. By matching students based on their propensity to enter FIGs

(as explained in the Methods section), we attempt to account for potential confounders as well as self-selection. After matching students, we then compare academic outcomes across the FIG and non-FIG students. While previous studies on freshmen seminars attempted to correct for selection bias via individual covariates, they only included a limited set of covariates, which may not sufficiently account for self-selection [42]. In our study, we match FIG and non-FIG students using nearly 200 covariates on a very large sample of students. Beyond this, to further delve into which specific aspects of a FIG are impactful to students, we also employ qualitative and text-mining methods on a large set of open-ended survey responses from FIG students. We believe this added analysis of first-hand accounts of the FIG program's effectiveness helps elucidate why freshman seminars tend to provide measurable benefits to student academic outcomes at a scale not previously found in the literature.

5.5 Methods

We present the methods for this project in the following manner. First, we give an overview of the data used in this study. Then, we give relevant definitions with respect to graduation and re-enrollment. Next, we give an overview of the methodology used in propensity score matching. Lastly, we discuss our approach to analyze student survey responses. The UW's Institutional Review Board approved all data collection and study design for this project.

5.5.1 Data

Data for this study spanned three distinct datasets: student registrar data, student survey data, and non-institutional data. We describe each in greater detail below. The information on students prior to their entry to the University from all three data sets ultimately served as a rich set of covariates that we used to predict the propensity for students to enroll in FIGs during their freshman year.

Student registrar data.

We collected de-identified student data from the University of Washington's (UW's) data custodians in early 2017. This data included complete student transcript records (courses taken, grades, etc.), student demographic information (race, gender, ethnicity, etc.), and student entrance application information (high school grades, entrance exam scores, etc.). We cleaned, curated, and stored this data in a secure SQL database on a secured server. We limited the data for this project to first-time, first-year students (freshmen) who first enrolled at the UW between 1998 to 2010. We used the year 2010 as a cutoff to allow at least 6 full calendar years for students to graduate with a baccalaureate degree (per the definition of graduate outlined below) as the data extended through 2016. In all, this included 18 years of institutional data and information on 57,979 unique freshmen students. We defined "FIG students" as those who had completed the 2-credit FIG seminar course during their freshmen year, as indicated on their transcript records. We defined "non-FIG" students as those who did not complete the FIG seminar course during their freshmen year.

Student survey data.

The UW's FYP collected exit surveys from nearly every FIG participant from 2010-2015. This was 14,514 students in total and an average of about 2,419 students per year. These surveys asked students a wide range of questions regarding their FIG experiences and most responses are recorded as open-ended text. For this project, we examined answers to a single question in the survey: "What did you find most valuable about the FIG?" for which there were 12,539 non-blank responses (86.4%). Note that the timeframes for the student registrar data described above and the survey data do not overlap except for the year 2010. Nevertheless, we still pulled student demographic information from the registrar data and linked it to all survey respondents, including students who were not included when analyzing the transcript data.

Non-institutional data.

In addition to the data from the University, we also used two additional sources of information: the College Board's enrollment planning services (EPS) data[1] and U.S. Census data. The EPS is an analysis and reporting service from the College Board which contains detailed information for almost every high school in the U.S. From the EPS, we used data on students' major intentions in higher education, students' preferences with regards to post-secondary campus settings (both with respect to the campus itself and the city in which it is located), students' long-term educational attainment goals, parents' educational attainment, and parents' income levels. EPS data was not available for individuals but was instead aggregated by high school. Accordingly, we aligned the EPS data to students' registrar data using their high schools from their applications to the University. We also used 2015 U.S. Census data on average income, average bachelor's degree attainment, and average high school completion for each ZIP code in the US. We aligned the Census data to each individual student using the ZIP code from their application to the University.

5.5.2 Defining graduation and re-enrollment

We labelled students as "graduates" if they completed at least one baccalaureate degree within 6 calendar years of first enrolling at the University. We determined this by looking at the difference between the term for which students' first baccalaureate degree was awarded and their respective first term on campus. We verified overall graduation rates with those used by the University's institutional reporting offices. We labelled students as "re-enrollments" if they returned for a 2nd academic year within 1 calendar year of the completion of their first academic year (i.e. they returned for their sophomore year within a year of completing their freshmen year). It should be noted that due to the large sample sizes involved in this study, we did not include any statistical tests of significance between

[1] See: `https://collegeboardsearch.collegeboard.org/pastudentsrch/support/licensing/college-board-search-services/enrollment-planning-service`

graduation and re-enrollment rates for FIG and non-FIG students. These tests would inevitably indicate even minute differences between the groups to be statistically significant due to large sample sizes. Instead, we focus on the practical significance of the differences.

5.5.3 Propensity score matching

As mentioned above, students selecting whether they will or will not enter a FIG presents the issue of selection bias and possible confounding [140]. In addition, factors that are predictive of participating in a FIG may also be associated with students' attrition or graduation. For example, a student whose parents have relatively low educational attainment may feel more unprepared for college and thus be more willing to seek additional help in the form of a FIG. At the same time, this students' parents' educational attainment may also be associated with lower academic success for the student. The existence of such factors makes it difficult to isolate the effect of FIGs from other possible factors that could affect the student outcomes of interest (in this case, graduation and re-enrollment).

As such, rather than simply comparing outcomes from treatment and control groups, PSM attempts to account for covariates that may influence the likelihood of receiving treatment (with "treatment" in this case defined as enrolling in a FIG). In PSM, a model is first constructed from potential confounders with a dependent variable associated with the treatment. The likelihood for each participant to receive treatment is referred to as the "propensity score." Participants in the treatment group are matched to those in the control group based on their propensity scores using a range of criteria/restrictions.

We used students' demographic information and pre-college information from the registrar data as well as information on their high schools from the EPS data and information on ZIP codes from the census data to calculate the propensity scores via a logit model. In total, this included 197 covariates, broken down as follow: 41 from the student registrar databases, 3 from the U.S. Census data, and 153 from the EPS data. It should be noted that this approach follows suggestions by Rubin and Thomas regarding the inclusion of all potential covariates in a propensity score model unless there is consensus that it is unrelated to the

outcome variable [142].

After calculating propensity scores, we matched students in the treatment (FIG) group to those in the control (non-FIG) group using two-levels of stratification and fixed caliper widths. We also employed different matching strategies when comparing the students, as described below. In terms of stratification, we first matched the students according to year of entry to the University and then by whether or not they were a STEM-interested student (i.e. a student interested in science, technology, engineering, and/or math (STEM) fields). We used the first stratum to better account for any institutional variation at the University across time. We used the second stratum to account for course difficulty as introductory STEM classes are considered amongst the most challenging for new students and these students may feel differential levels of engagement while taking with these introductory classes [70]. STEM interest was determined by whether students' major declaration during their first term was in a STEM field, whether they had a pre-major designation that was associated with a STEM field during their first term, and/or whether they took any STEM gatekeeper classes during their first term[2]. We coded STEM interest in a binary manner. Using this dual stratification, we matched every FIG student to corresponding non-FIG students from the same entrance year and with the same (binary) indication of STEM interest.

After the above stratification, we matched students based on caliper matching, wherein students were matched if they had propensity scores within a specified interval from each other. We kept the caliper at one-tenth of the pooled standard deviation of all propensity scores. This is half the caliper width recommended by Austin [16], thereby giving more stringent matching. Using the caliper, we matched students in each group in three different ways: one to many (where the results from the matched control group were averaged), one to one with replacement (where each treatment student is only matched to a single control student, but not vice versa), and one to one without replacement (where each treatment student is only matched to a single control student and vice versa). After matching students,

[2]STEM gatekeepers are introductory STEM classes that serve as prerequisites for advanced STEM classes. They are also key determinants of whether students are accepted to highly-competitive STEM majors

we compared the graduation and re-enrollment rates of the student groups.

We also performed an additional round of matching to better control for course difficulty. In this round of matching, we only included matches if the FIG and non-FIG students had completely identical coursework for the academic term in which the FIG student completed the FIG seminar. We only included courses that were completed for numeric grades (i.e. not pass/fail) in this matching and the FIG seminar course was excluded when finding students with identical coursework. After matching, we compared students' grade point averages (GPAs) for the term where students had identical coursework in addition to comparing their eventual re-enrollment and graduation.

In addition, we also compared Hispanic and under-represented students based on FIG entry. When examining these student subgroups, we only included FIG and non-FIG students from the specific student subgroup for matching and all other students were excluded. As with all students, we used a dual stratification across year of entry and STEM interest and the caliper widths were reset based on one-tenth the pooled standard deviation for each student subgroup.

5.5.4 Survey analysis

To code survey responses describing what students found most valuable about FIGs, we relied on a topic modeling-based qualitative analysis approach using grounded theory. First, we developed an initial codebook using a form of topic modeling called Latent Dirichlet Allocation (LDA) [27]. LDA is a generative statistical model that allows for topic discovery. In LDA, "documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words" [27]. In the case of the survey responses, the documents refer to individual student responses and the topics are general themes in the survey responses as represented by associations with specific words. Prior to generating the topic models, we removed stop words from the survey responses and stemmed the remaining words. We generated five initial topics using LDA. This allowed for the distinction of preliminary tags/codes of interest and we only used topic modeling as a starting point in

coding the massive text corpus.

From there, two researchers independently coded an initial set of 1000 survey responses and iteratively developed a joint codebook to use. The researchers used each tag within the codebook in an independent manner for each response unless specifically noted in the Results section (i.e. the use of one tag did not exclude the use of another). There was also no limit to the number of tags that could be applied to each response. After coding the initial 1000 responses together, the researchers each coded an additional 1000 responses independently and discussed consistency in coding thereafter. Then, the researchers coded every remaining response individually. After coding all 12,539 responses individually, each response that was not coded identically between the two researchers was discussed and a consensus regarding coding was drawn. The researchers were only provided with the text of the student survey responses and the codebook when coding responses; student demographics and academics were not visible in any way. We used the final set of tags for each individual student response in the analysis.

5.6 Results and Discussion

5.6.1 FIG participation

Of the 57,979 students in the study, 32,572 enrolled in a FIG (56.2%) while 15,407 of them did not (43.8%). Table 5.1 shows the demographics of the FIG and non-FIG students. There were a few relatively large differences in demographic composition among FIG and non-FIG students with respect to gender and race. In particular, female and Caucasian students were over-represented among FIG students while male and Asian students were under-represented among FIG students. For the most part, all other demographics were fairly consistent in terms of proportions across the two groups. Interestingly, Tokuno reported that in the early years of the FIG program at UW, African American, American Indian, and Hispanic students were less likely to enroll in FIGs than other demographic groups [181]. This, however, was not the case with the students examined in this study as each of these groups were slightly

more represented among FIG students.

Table 5.1: Demographic Overview of FIG and non-FIG Students

	FIG Counts (%)	non-FIG Counts (%)
Total	32,572	25,407
Gender		
Female	18,494 (56.78%)	12,146 (47.81%)
Male	14,049 (43.13%)	13,233 (52.08%)
Unidentified	29 (0.09%)	28 (0.11%)
Race		
African Am.	1,026 (3.15%)	733 (2.89%)
Am. Indian	504 (1.55%)	327 (1.29%)
Asian	8,610 (26.43%)	8,144 (32.05%)
Caucasian	19,439 (59.68%)	13,172 (51.84%)
Haw./Pacific Is.	274 (0.84%)	162 (0.64%)
Unidentified	2,719 (8.35%)	2,869 (11.29%)
Ethnicity		
Hispanic	1,699 (5.22%)	1,158 (4.56%)
Not Hispanic	30,873 (94.78%)	24,249 (95.44%)
Residency		
Residents	27,540 (84.55%)	21,749 (85.60%)
Non-Residents	5,032 (15.45%)	3,658 (14.40%)

The number and proportion of freshmen who completed FIGs across time is shown in Figure 5.1. Both increased steadily from 1998 through 2003, only to level off thereafter. In terms of percentages, FIG participation increased steadily from 40.2% in 1998 to 66.2% in 2003, only to remain fairly level from there and never dropping below 55%.

5.6.2 Propensity score matching

The distributions of the propensity scores for FIG students, non-FIG students, and non-FIG students matched to FIG students in a one-to-many manner are shown in Figure 5.2. The mean (±SD) propensity score for FIG students was 0.61 (±0.13) and the mean (±SD)

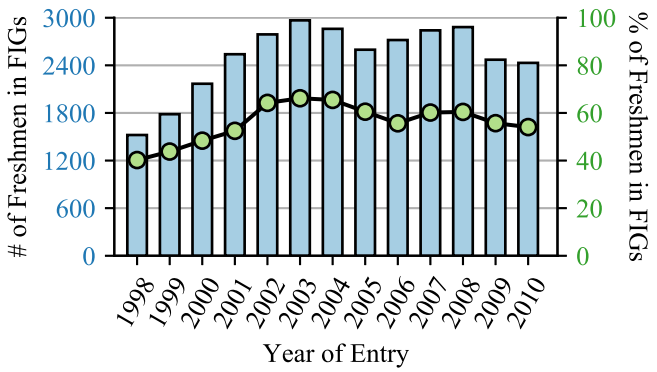

Figure 5.1: FIG enrollment as count of freshmen (bar chart) and percentage of freshmen (line chart)

propensity score for non-FIG students was 0.50 (±0.18), where, as noted earlier, these propensities represent the probability of a student enrolling in a FIG, regardless of whether they ultimately did. As such, the fact that FIG students had higher propensities than non-FIG students should come as no surprise. The shape and range of the distributions of propensity scores across FIG and non-FIG students show a large amount of common support, thereby indicating a high number of potential matches across FIG and non-FIG students. The caliper used when matching all students was 0.0163 - each FIG student was matched to corresponding non-FIG students who were within ±1.63% as likely to enroll in a FIG. The resulting distribution of matched non-FIG students consists of the average propensity of all non-FIG students who fall within the designated caliper for each FIG student. The shape of the distributions of propensity scores of FIG and matched non-FIG students is nearly identical. The distribution of matched non-FIG students had a mean propensity score of 0.61 (±0.13), which was identical to the mean and standard deviation of the propensity score of FIG students.

To evaluate the effectiveness of the PSM, we calculated the standardized bias across each of the variables used in the PSM. The standardized bias calculation, as described by Caliendo, is a "measure to assess the marginal distance of the (variables)" used in the PSM

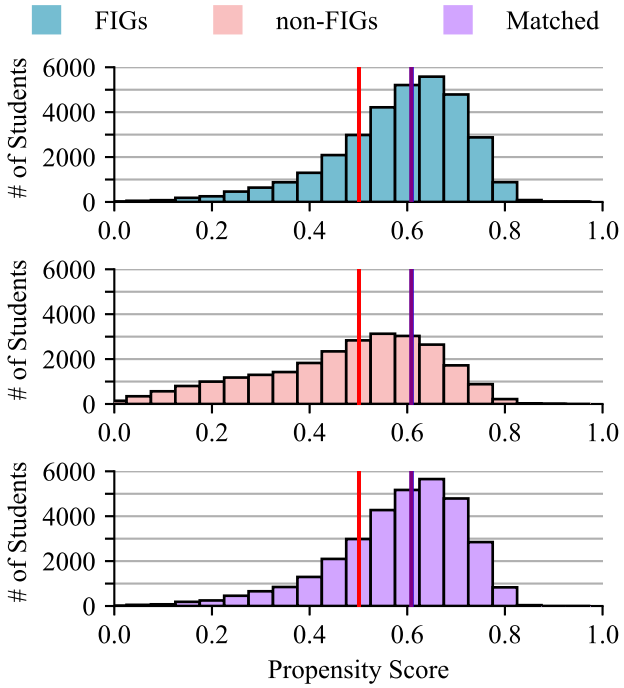

Figure 5.2: Propensity score distributions for FIG students (top), non-FIG students (middle), and matched non-FIG students (bottom). The top and bottom distributions were used in the analysis. The left and right vertical lines across each distribution indicate the mean propensity score values for non-FIG students (0.50; from the middle distribution) and matched non-FIG students (0.61; from the bottom distribution), respectively. The mean of the FIG students was approximately equal to that of matched non-FIG students.

[36]. In most empirical studies, a bias reduction below 3-5% is typically seen as sufficient in reducing potential confounding [36]. After PSM, the standardized bias across all variables used in PSM had a mean value of 1.3% and a median value of 1.1%. Only 3 of the 197 variables had a standardized bias value greater than 5% and none were greater than 6%.

When looking at one-to-many matching across all students, 32,512 FIG students (equivalent to 99.8% of the FIG population) had at least one non-FIG student matched. Of these matched students, 30,257 FIG students (93.0%) had at least 20 non-FIG students matched, again indicating a high level of common support for PSM. Each FIG student was matched to an average (±SD) of 66.1 (±34.0) non-FIG students and the distribution of matches per

FIG student is shown in the top of Figure 5.3, with colors indicating the quartile of the propensity score of the FIG student. As can be expected, students with propensity scores in the second and third quartiles had the greatest number of matches while FIG students who were more on the fringes of the propensity score distribution tended to have fewer matches. The bottom of Figure 5.3 shows a cumulative frequency graph of the number of matches for FIG students.

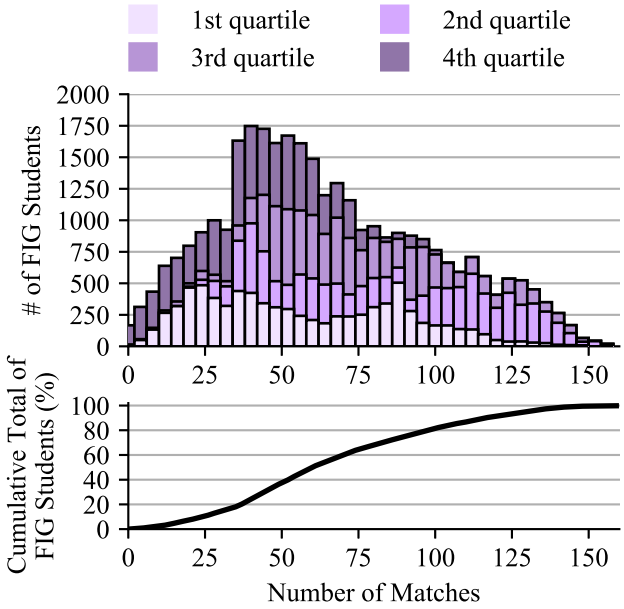

Figure 5.3: Distribution of the number of matches for each FIG student when using PSM (top). Colors indicate FIG student propensity score split by quartiles. Cumulative frequency graph of matches for FIG students (bottom). Line indicates cumulative percentage of students who have at most the corresponding number of matches.

5.6.3 Retention and graduation rates

Unadjusted rates.

Prior to examining the differences between FIG and non-FIG students using PSM, we calculated unadjusted overall graduation and re-enrollment rates. Graduation and re-enrollment

rates for the entire population were 78.6% and 92.7%, respectively. Graduation and re-enrollment rates for the FIG students were 81.6% and 94.2%, respectively; graduation and re-enrollment rates for the non-FIG students were 74.9% and 90.7%, respectively. FIG students had vastly higher unadjusted graduation and re-enrollment rates than their non-FIG peers and, across the University as a whole, rates were substantially higher than national averages [109]. Rates for all FIG and non-FIG students across time are shown in Figure 5.4. For every entry year examined, we found that FIG students had higher graduation and re-enrollment rates than their non-FIG counterparts. Note that these rates were prior to matching students using PSM.

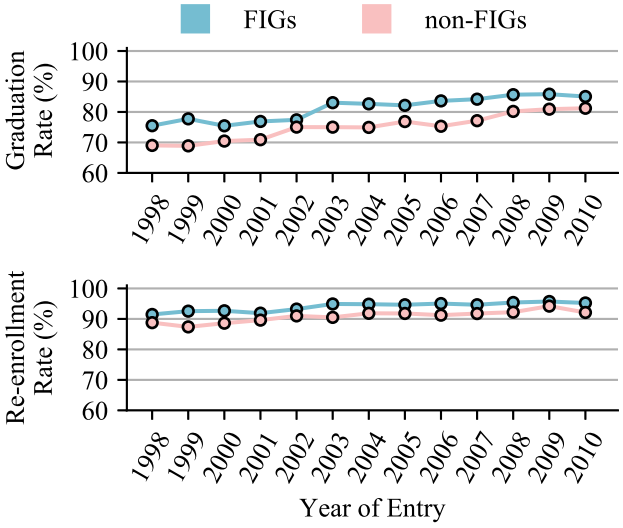

Figure 5.4: Graduation rates (top) and re-enrollment rates (bottom) for FIG and non-FIG students across time.

Rates After PSM.

Graduation and re-enrollment rates for FIG and non-FIG students after PSM are shown in Table 5.2. The results did not significantly change based on the matching strategy used and only results from one-to-many matching are discussed in further detail. After matching,

FIG students tended to have substantially higher graduation and re-enrollment rates than their matched non-FIG peers (differences of 7.0 and 3.7 percentage points, respectively). These differences amount to decreasing the overall institutional attrition rate by 27.5% and the institutional first-year attrition rate by 38.8%. The percentage point differences in re-enrollment between FIG and non-FIG students are greater than those found by Tokuno (1.6 percentage points higher for FIG students) [181] but lower than those found by Tinto (over 6 percentage points) [178] at the same university. Additionally, these differences in re-enrollment rates are lower than those noted in previous studies comparing freshman seminar participants with non-participants (e.g. [149, 113, 166]). It should be noted, however, that for both groups, the re-enrollment rates are greater than 90%, thereby decreasing the potential maximum difference between the two groups in terms of percentage points. While these differences in re-enrollment rate percentage points are in line with the PSM-based study conducted by Clark, the differences between baseline rates are much less pronounced in this case [42].

Table 5.2: Graduation and re-enrollment rates for all students after PSM. Reduction refers to the difference in FIG and non-FIG rates divided by the non-FIG attrition rate ($\frac{\text{FIG rate - non-FIG rate}}{\text{100\%- non-FIG rate}}$)

Matching	Measure	FIG	non-FIG	Difference	Reduction
One-to-many,	Graduation	81.60%	74.57%	+7.03%	27.6%
w/ replacement	Re-Enrollment	94.18%	90.49%	+3.69%	38.8%
One-to-one,	Graduation	81.60%	74.02%	+7.58%	29.2%
w/ replacement	Re-Enrollment	94.18%	90.04%	+4.14%	41.6%
One-to-one,	Graduation	81.51%	74.84%	+6.67%	26.5%
w/o replacement	Re-Enrollment	94.18%	91.05%	+3.13%	35.0%
One-to-many,	Graduation	84.02%	73.82%	+10.20%	39.0%
w/ replacement &	Re-Enrollment	94.90%	92.00%	+2.90%	36.3%
matched courses	GPA	3.214	3.011	+0.203	

Matching Coursework.

When matching students only if they had identical coursework during the term in which the FIG student completed the FIG seminar, 2,078 students were matched. Of these, only 692 (about one-third) were matched to more than one student. The graduation rates, re-enrollment rates, and GPAs for FIG and non-FIG students after PSM are also shown in Table 5.2. Controlling for coursework, FIG students had substantially higher GPAs (+0.2 on a 4.0 scale) when taking the same courses as their non-FIG counterparts. Interestingly, this analysis did not account for which courses were specifically part of the FIG and only looked to see that the students were taking the same courses. For example, this allows for FIG students who took 3 courses as part of their FIG (the FIG seminar excluded) to be compared to non-FIG students who took the same 3 courses. It could also allow for FIG students who took the FIG seminar as a stand-alone course alongside 3 additional courses to be compared to non-FIG students who took the same 3 courses. Regardless, the results here indicate that FIGs tend to have a substantial effect on student academic success (as measured by GPA) regardless of whether the students take courses as part of the FIG cluster or not. The specific benefits of taking courses as part of a FIG versus not were not examined in this analysis. In addition to this difference in GPA, the difference in graduation and re-enrollment rates for FIG and non-FIG students follow the same trend as above, though to different degrees. Using coursework in addition to PSM to match students yielded a difference in graduation rate that was greater than that noted above (10.20 percentage points) while the difference in re-enrollment rates was smaller than that noted above (2.90 percentage points).

Examining specific student groups.

In addition to examining the graduation rates for all students, specific student groups were also examined. We compared students who self-identified as being either Hispanic or of an under-represented minority group across FIG entry. The list of under-represented minority groups is maintained by the UW's Office of Minority Affairs and Diversity and includes

federally-recognized underrepresented minority students such as African American, American Indian/Alaska Native, Hawaiian/Pacific Islander, and Latino students. As was done with all students, we matched students who enrolled in a FIG with non-FIG students using PSM while stratifying on year of entry to the University and STEM intention. We reset the caliper width for each subgroup (i.e. Hispanic and under-represented students) and a one-to-many matching strategy was employed. The caliper when matching Hispanic students was ±1.56% and for under-represented students was ±1.52%. Of the 1,699 Hispanic students in FIGs, 1,546 (91.0%) were matched to at least one Hispanic non-FIG student; of the 1,578 under-represented minority students in FIGs, 1,428 (90.5%) were matched to at least one under-represented non-FIG student.

Table 5.3: Graduation and re-enrollment rates for Hispanic and under-represented students after PSM. Reduction refers to the difference in FIG and non-FIG rates divided by the non-FIG attrition rate ($\frac{\text{FIG rate - non-FIG rate}}{\text{100\%- non-FIG rate}}$)

Subgroup	Measure	FIG	non-FIG	Difference	Reduction
Hispanic	Graduation	77.96%	69.32%	+8.64%	28.16%
	Re-Enrollment	93.06%	87.28%	+5.78%	45.44%
Under-Represented	Graduation	76.81%	62.81%	+14.00%	37.64%
	Re-Enrollment	94.75%	88.18%	+6.57%	55.58%

The Graduation and re-enrollment rates after PSM across student subgroups are shown in Table 5.3. A few things are of note when looking at these rates. First, the graduation and re-enrollment rates for Hispanic and under-represented FIG students were substantially higher than their non-FIG peers. This difference is much more pronounced than the estimated FIG effects for all students in Table 5.2, except when matching on coursework. Second, the graduation rates for both Hispanic and under-represented students were still below the University's average across all students. Third, the re-enrollment rates for Hispanic and under-represented students were also below the University's average except for under-represented FIG participants. These facts could go hand-in-hand as the lower average

rates across these groups allows for more differential gains in terms of percentage points to be realized for students attending FIGs. We know that the FIG curriculum at the University is not tailored to specific student groups based on race, ethnicity, or family backgrounds. We intend to further examine this greater effect of FIGs on Hispanic and under-represented students in future work, as well as potentially looking at first-year college entrants.

It becomes apparent when examining the results from PSM across Tables 5.2 and 5.3 that the observed differences in FIG and non-FIG students are very robust to different matching strategies and across student subgroups. Regardless of the matching strategy employed, differences between the graduation rates for FIG and non-FIG students were always at least 6.67 percentage points while differences in re-enrollment rates were always at least 2.9 percentage points. The fact that the PSM graduation and re-enrollment rates were similar to the unadjusted graduation and re-enrollment rates highlighted in Figure 5.4 is in part due to the number of possible matches based on FIG students' propensity scores, as nearly the entire FIG population was matched to at least one non-FIG student. This also speaks to the degree to which there was common support amongst the propensities of FIG and non-FIG students, without which the given number of matches would not be possible. This wide common support is indicative that the differences between FIG and non-FIG students, as described by the variables used in the propensity score model, were not extreme. It should also be noted that even when matching students on more stringent criteria (i.e. matching students with identical coursework), far fewer students were matched and the differences between FIG and non-FIG students still persisted.

5.6.4 Survey analysis

The above quantitative analysis sheds light on the degree to which FIGs impact student outcomes. To better understand specific aspects of FIGs that students found to contribute positively to their education experience, we analyzed student survey responses using an approach that relied on both topic modeling and grounded theory.

Developing a codebook.

We used topic modeling (namely, LDA) to develop an initial set of tags to use when examining responses from the student surveys. We provide examples of student responses (and associated tags) in the appendix. We used five topics in the LDA model and ten words associated with each of the topics are shown in Table 5.4 with three common terms (FIG, valuable, part) excluded from each word list. Apparent from the topic/word list is that students tended to talk about their FIG leader and/or the FIG program, information on majors, information on campus resources, meeting people and/or making friends, forming study groups in a class, learning about UW, and learning about the city of Seattle.

Table 5.4: Topic modeling results. Words are listed in descending order of probability to appear in a topic

Topic	Words
FIG leader and/or program and/or cluster	class, leader, really, classes, quarter, made, program, us, time, helpful
information on majors and/or resources	learning, UW, valuable, majors, major, resources, different, information, future, program
meeting people and/or making friends	people, meeting, friends, new, meet, making, met, interests, similar, study
forming study groups	people, classes, group, study, students, class, able, know, could, meeting
learning about UW and/or Seattle	Seattle, know, people, community, UW, get, college, project, school, communities

Using topics from LDA, we eventually developed the codebook detailed in Table 5.5. The codebook consisted of 18 unique tags and included a *DROP* tag that was used for responses

that were deemed irrelevant to the survey question. These responses were subsequently removed from the analysis and any percentage counts. The tags examined the responses from the perspective of academics (e.g. the *finding interests* and *getting into classes* tags), social integration (e.g. the *meeting people*, *community*, and *people in classes* tags), program organization (e.g. the *cluster* and *connected classes* tags), seminar design (e.g. the *activity*, *smaller class*, and *course* tags), and general college transition (e.g. the *transition* and *skills* tags). As noted previously, all tags were applied in an independent manner (that is, the use of one tag did not exclude the use of another) unless in the cases of the *interests* tag, which was only used in conjunction with the *meeting people* tag.

Coding results.

Two coders independently coded each of the 12,539 survey responses, with tags applied independently and in a mutually exclusive manner. There was also no limit to the number of tags that could be applied to a single response. When examining the responses of all students, 7 tags were used in at least 5% of all responses: *new people* (52.4%), *people in classes* (23.8%), *activity* (13.7%), *survival skills* (13.4%), *FIG leader* (9.2%), *interests* (6.0%; used in conjunction with *new people*), and *transition* (5.3%). All other tags were used in less than 3.5% of responses and were not analyzed further. The frequencies with which the 7 tags were applied are shown in Table 5.6.

In all, the two coders coded 8,897 (71.0%) of all responses identically, regardless of the number of tags applied to each response. Table 5.6 also shows the degree to which the coders applied similar tags on a tag-by-tag basis, both in terms of general percent similarity as well as Cohen's kappa values calculated for each tag independently. Metrics were not calculated for the *similar interests* tag as it was only used in conjunction with the *meeting people* tag. In all, the coders had at least 90.7% agreement across each of the 18 tags. Additionally, each of the 6 tags that were used in at least 5.0% of responses and analyzed had a Cohen's kappa value of at least 0.66, with 3 of the 6 having Cohen's kappa values greater than 0.85. These metrics were calculated before the coders discussed discrepancies in their coding and arrived

Table 5.5: Codebook used in analysis of survey responses

Tag	Description of what tag refers to
activity	an activity that occurred in the FIG seminar
cluster	the FIG course cluster and easier registration for courses
community	community and camaraderie among FIG students, especially with respect to shared experiences
college goals	helping students outline their goals for college and their long-term academic aspirations
connected classes	overlapping topics/themes across FIG classes
course	some non-activity aspect of the FIG seminar and/or the general design of the seminar
FIG leader	the FIG leader(s) and peer mentorship from an upperclassman
finding interests	helping students explore areas of study (not majors)
getting into classes	using the FIG to get into reserved course sections
majors	helping students learn about majors available on campus
new people	meeting new people, making friends, and/or forming a social group of some kind
none	nothing
people in classes	having familiar people in classes and having people to study with
sharing interests	meeting people with common interests (only used in conjunction with *meeting people*)
survival skills	general campus survival skills and learning about resources available on campus
smaller class	having the FIG seminar be a smaller class compared to other classes
transition	helping students with the shift from high school to college
DROP	exclude response from analysis (applied to 292 (2.4%) of responses)

at a consensus across all tags. All calculations regarding the frequency of applied tags are based on tags after the coders came to a consensus regarding non-identical tags.

Far and away, students thought that social aspects of the FIG were the most valuable with over half of all responses referencing meeting new people and/or making friends. Meanwhile, the second most frequently applied tag referenced knowing people in classes and being able to form study groups. Also interesting is that about one-tenth of those mentioning meeting new people also voluntarily shared that meeting people who share common interests with them was also important. This idea of greater social integration within a campus, be it by making friends and/or meeting people, is frequently visited in retention theory as a factor in increasing retention [175]. In this case, over 70% of all respondents mentioned some social impact the FIG had on their first term (across the *new people*, *people in classes*, *FIG leader*, and *community* tags). The idea of forming study groups in classes also ties into the idea of first-year students being academically supported as they adjust to college-level study and gain confidence [133]. This points back to Barefoot's definition of freshman seminars and how they "provide essential skills for college success" and also can be integral to them forming a peer group [22].

The third-most applied tag referenced survival skills on campus. Most often, this tag was used in reference to students understanding how to succeed in college, such as how to use library resources, how to register for classes, and where to find help with homework. The fourth-most applied tag referenced activities that were part of the seminar, frequently referring to an activity wherein the FIG leaders led their students off campus to explore the city. Most students referenced this activity in the context of being able to better acquaint with their peers outside of a class while also learning more about the city they are living in. It should be noted that as the largest post-secondary institution in Washington State, the UW enrolls many students who have never lived in the city of Seattle and/or in an urban setting. The FIG students also frequently mentioned their FIG leader as being a valuable asset. Much of this was in reference to having an undergraduate peer who had been in their position as a freshman and knew how to navigate the University, both physically and academically. In some cases, students also referenced their FIG leader as being an effective sounding board with regards to struggles as they adjusted to college life.

Interestingly, when examining student responses across subgroups, we did not observe salient differences in the frequency of applied tags as shown in Table 5.6. Each of the tags that were applied to at least 5% of the responses across all students (i.e. the 7 tags listed in Table 5.6) were applied with about the same frequency for both Hispanic students and non-Hispanic students ($\pm$1.3 percentage points). When looking at the same for students from under-represented minority groups, only references to meeting new people and activities in the FIG seminar differed by more than $\pm$1 percentage points from students not from under-represented minority groups. More specifically, the tag referencing meeting new people was used in 50.4% of responses for under-represented students and 52.5% of responses for non-under-represented students (2.1 percentage point difference) while the tag referencing a specific activity in the FIG was used in 16.1% of responses for under-represented students and 13.5% of responses for non-under-represented students (2.6 percentage point difference). This suggests that under-represented students could either find it less valuable to meet new people and make friends via FIGs than their peers or just have a more difficult time doing so. Regardless, the observed differences in the frequency of applied tags are still rather small across all groups (<3%).

It does not seem as though specific student subgroups found the FIGs more valuable than their peers in any particular way based on their subjective responses. This is particularly interesting when considering the larger impact FIGs had on these groups in terms of academic success, as shown in Table 5.3. One potential explanation of this is that FIGs these students are more unprepared for their transition to college that the FIGs assist with, albeit in a manner that is not differentiable across student groups. An alternative explanation is that FIGs do have some differential impact on these groups with respect to the tags examined but it is not subjectively noted or articulated by the student subgroups. We intend to examine both possible explanations in greater detail in the future.

In all, the results from this analysis of student survey responses help provide further context to the quantitative results from PSM in understanding why FIGs have the observed impacts on student academics. Social aspects of the FIGs tend to be the primary positive

takeaway for students and this includes meeting peers, finding groups/peers to study with, and having a peer mentor in the FIG leader. Beyond that, students also found specific activities within FIGs to be useful as well as introductions to different resources available on campus to students. These two ideas - a greater social integration with the campus as well as providing necessary academic support - are not only central tenants to many freshman seminars, but they also hold a prominent place in longstanding student attrition theory [57].

5.6.5 Limitations

We understand that there are several limitations with this study. First, although using PSM explicitly accounts for observable differences between FIG and non-FIG students before treatment, our analysis only balances the means of those observed covariates between the treatment and control groups. Thus, the results of this study may remain subject to biases of unobserved confounding variables [97]. In the absence of a randomized controlled trial, we believe PSM allows for a more robust analysis than simply comparing FIG and non-FIG groups without any matching and, in this case, we match students while explicitly adjusting for selection bias and confounding variables using nearly 200 observables. What's more, our study provides an example of using matched comparisons with large-scale and detailed data to gauge freshmen seminars, which has rarely been done in prior work.

Another limitation is that this study relies on data from a single institution in its analysis. This limits the degree to which the results can be generalized to other campus settings. We also understand there are institution-specific subtleties with the data that may not be apparent when comparing to other institutions' data. The data was limited to a single university due to the difficulty in obtaining detailed, anonymized registrar records in higher education. This type of data is different than longitudinal data typically used in education studies as it centers on course-level transcript records and individual-level demographic information, which is protected by institutions and not readily available for research purposes.

Another limitation of this work is the potential threat of interpreting FIG effects in light of other concurrent events. For instance, financial aid and on-campus residency may be

other mechanisms that may differ among students concurrently with FIGs and may drive the differences observed between FIG and non-FIG students. As examples, previous studies have examined how financial aid can impact student persistence (e.g. [34] and [171]) and other studies have looked at the effect of on-campus residency on persistence, even among first-year students (e.g. [134]). However, this data was not available through the University's data stewards and therefore, we could not gauge the degree to which the estimated FIG effects were robust with respect to these possible alternative mechanisms.

A further limitation of this study is the fact that the surveys did not align with the data used in PSM. This was due to the timeframe used to define successful graduation (i.e. 6 years), setting students first enrolling in 2010 as the last cohort examined. UW's FYP, meanwhile, did not electronically collect student feedback on FIGs until 2010. Despite this limitation, we believe it is still useful for us to provide this information. As aforementioned, the survey results corroborate the PSM estimates of FIG effects as students generally found FIGs useful. The aspects of usefulness of FIGs perceived by students are also predictive to student academic outcomes, particularly with respect to social integration, as informed by prior work (e.g. [177]). The quantitative and qualitative results are related in that they both shed light on the FIG program, but provide different insights across distinct data sources.

5.6.6 Directions for Future Research

We believe a next step for this work is to examine the degree to which different FIG course compositions impact student success. More specifically, because FIGs at the UW allow for students to take varying numbers of credits, we hope to center our analysis on treating FIG participation as a non-binary treatment effect based on the number of credit hours students take as part of the FIG. This would then allow for a more granular analysis with respect to FIG treatment effects and outcomes as one could hypothesize based on the results shown in this work that a greater involvement in a FIG (i.e. more credit hours as part of a FIG cluster) leads to greater student success.

We believe another area for future research is to examine the degree to which student

success relates to characteristics of their FIG leaders. FIG leaders must take a preparation course before leading a FIG, thereby allowing us to identify these students based on their transcript records. We can then leverage the student registrar data to look at the academics of FIG peer instructors to see if specific attributes of the leaders relate to more effective administration of the seminar course. We believe this work can have wide implications in the design of future freshman seminars and the hiring of potential peer mentors for these seminars.

Lastly, we believe a third space for more research involves better understanding why FIGs had a greater effect on the academic success of Hispanic and under-represented minority groups. We intend to examine the possibility of an ethnographic exploration of how these students go through the freshman seminar and how it influences their first year. This may also be expanded to examine first-generation students, which were not identified in this work. Additionally, we also intend to examine FIG cohort composition and the degree to which homophily impacted student success in some way. This includes examining specific minority groups in select fields (e.g. women in STEM) and seeing if their success was impacted by having a similar peer makeup in their FIG seminar. An additional route of exploration may involve examining student transitions across fields, be it in switching majors of interest or showing some tendency to do so, as we have examined in the past with other student groups [15, 7].

5.7 Conclusions

This work examined the impact of FIGs on undergraduate student outcomes using both quantitative and qualitative approaches. Using student registrar data and propensity score matching, we show that FIGs positively impact both graduation and re-enrollment rates for participants, an effect that is more pronounced for Hispanic students and students from under-represented minority groups. We also show evidence that FIGs tend to positively impact academic performance as measured by GPA. Then, using open-ended survey responses from FIG participants, we tease out what FIG participants found most valuable about the

program, which centered around social aspects of the program (namely making new friends and knowing people in classes). Other aspects of interest included specific activities within the FIG seminar, general campus survival skills, and having an undergraduate peer leading the seminar course.

5.8 Chapter Appendix

Examples of student responses.

Below are examples of student survey responses as well as the corresponding tags given to them by the researchers.

- *"The most valuable part of the FIG to me was being with a group of people who were taking the same classes as myself. This helped a lot with finding study groups and people I get along with."* - people in classes
- *"I found the FIG program a waste of time and did not see the importance of it. My FIG was confusing and not helpful in my transition into UW. It took up my time to work on my other classes."* - none
- *"I enjoyed getting to know my FIG teacher because she knows so much about UW and is able to give me guidence* (sic) *on issues when no one else I know would have been able to help."* - fig leader
- *"Having all my classes with the same 24 kids and also having a class with just 24 kids. This made the transition into college less overwhelming and easier. It was also very nice to get to know kids on campus that I would have otherwise never met."* - new people, people in classes, transition
- *"The opportunity to get to know people with the same interests as me, because sometimes it can be very hard to find people you can relate to."* - new people, interests
- *"Getting introduced to the UW because I felt pretty overwhelmed coming in, so it was pretty helpful."* - skills
- *"Getting to know the other students in my FIG was by far the most valuable component*

of the experience. It was so nice to have the exact same classes and assignments as so many people, it made the support system feel much stronger." - new people, people in classes

5.9 Publications/Presentations

- Lovenoor Aulck, Joshua Malters, Casey Lee, Gianni Mancinelli, Min Sun, and Jevin West. Helping Students FIG-ure It Out: A large-scale study of freshmen interest groups (FIGs) and student success. *Research in Higher Education*, 2019 ***(journal paper - in draft)***
- Lovenoor Aulck, Joshua Malters, Casey Lee, Gianni Mancinelli, Min Sun, and Jevin West. Helping students FIG-ure it out: Using institutional records and student survey responses to examine freshmen interest groups (FIGS). *Conference of the Society for Research on Educational Effectiveness (SREE)*, 2019 ***(conference presentation)***

Table 5.6: Frequency of applied tags (only those in at least 5% of all responses are shown). Percent agreement refers to percentage of responses for which coders applied the tag in an identical manner. Percentages for students indicate percent of responses from student group that were tagged with respective code.

Tag	Percent Agreement	Cohen's Kappa (κ)	All students	Under-represented	Non-under-represented	Hispanic	Non-Hispanic
new people	92.9%	0.859	52.4%	50.4%	52.5%	53.5%	52.2%
people in classes	95.5%	0.871	23.8%	23.7%	23.7%	24.5%	23.6%
activity	90.7%	0.660	13.7%	16.1%	13.5%	13.1%	13.7%
survival skills	93.7%	0.666	13.4%	12.6%	13.5%	13.4%	13.4%
FIG leader	98.6%	0.914	9.2%	8.2%	9.2%	9.4%	9.1%
sharing interests	-	-	6.0%	6.0%	6.0%	6.3%	6.0%
transition	98.2%	0.786	5.3%	5.0%	5.3%	5.4%	5.3%

Chapter 6

CONCLUSIONS

Traditional universities are faced with existential crises on several fronts - from the increasing popularity of online degree programs and MOOCs to budgets that have never recovered from the Great Recession to continually scrutinized pricing policies in light of expanding student debt. With this in mind, it becomes imperative that these institutions maximize the utility of resources they have, including their institutional data, in improving their educational ecosystems and the lives of students. It can no longer be the case that institutional data is used for purposes of record-keeping alone; it must be explored, interrogated, and analyzed in a manner that institutionalizes a culture of data-informed decision-making. In an age where billion-dollar tech companies have numerous data science teams across their organizations, it is mind-boggling that schools with billion-dollar footprints have few, if any, data scientists on staff.

With respect to this, in the previous four Chapters, I demonstrated how data that is routinely collected by institutions of higher education can be leveraged to better understand students and educational ecosystems. The work I presented did not rely on additional data collection processes/procedures but instead focused on using data that already exists. In so doing, I share numerous findings while examining freshmen students prior to enrolling in college and during their first year. I briefly summarize the primary contributions of each of the four substantive research Chapters below.

- Chapter 2 examined the degree to which we can predict student enrollment and then used this information to allocate a scholarship fund to domestic non-resident students.
    1. This Chapter demonstrated that student enrollment can in fact be predicted using machine learning approaches

    2. This Chapter also demonstrated how numerical optimization techniques can be used to allocate scholarship awards in an effective manner that is flexible with respect to objectives
    3. This Chapter found some evidence that students of lower merit tend to be more receptive to scholarship awards than those of higher merit
- Chapter 3 examined the effects of scholarship awards on student enrollment.
    1. This Chapter leveraged the approach deployed in the previous chapter, demonstrating how machine-optimized award distributions can allow for clean experimental designs
    2. This Chapter, much like the one before, found some evidence that the effect of scholarships was more pronounced for students of lower merit
- Chapter 4 explored predicting student attrition from their first-year's data
    1. This Chapter found that student attrition can be accurately predicted given a limited amount of information from students' first years
    2. This Chapter also found that features derived from students' transcripts were much more useful for making these predictions than pre-entry or demographic features
- Chapter 5 analyzed the impact that a freshman seminar program had on student performance and persistence
    1. This Chapter found that these seminars positively affect student graduation, persistence, and performance
    2. This Chapter used first-hand responses from FIG students to better understand which aspects of the seminar were most beneficial, finding that students found social aspects to be the most useful

The work highlighted in this dissertation will continue on in many ways. Personally, this dissertation has laid the groundwork for my transition into a position within institutional research immediately after my doctorate is complete. This will see me working in a new role within the University of Washington's (UW's) administrative circles on using data science to

help administrators make more informed decisions. I envision this role as one that will grow as the UW continually learns how to best leverage its own data and administrators see value in data-centric tools and methods. In all, much of the momentum behind the role centers on the work in this dissertation.

My post-dissertation plans also speak to a larger change not only at the UW, but across institutional research. Where institutional research meeting agendas previously found no mention of data science approaches and techniques, now discussions are continually involving more terms like machine learning, artificial intelligence, and analytics. There is a growing curiosity on how these tools can be leveraged and implemented in institutional circles. What remains to be seen, however, is whether this change also signals a cultural shift amongst institutional policy makers in embracing data-centric approaches as a means to answer some of the most pressing challenges higher education is facing - a shift that I believe is much needed.

BIBLIOGRAPHY

[1] university 101 programs - history of the first university seminar & the university 101 program.

[2] WWC Intervention Report: First Year Experience Courses. Technical report, US Department of Education, 07 2016.

[3] Everaldo Aguiar, Nitesh V Chawla, Jay Brockman, G Alex Ambrose, and Victoria Goodrich. Engagement vs performance: using electronic portfolios to predict first semester engineering student retention. In *Proceedings of the 4th International Conference on Learning Analytics And Knowledge*, pages 103–112. ACM, 2014.

[4] Dennis Ahlburg, Michael McPherson, and Morton Owen Schapiro. Predicting higher education enrollment in the United States: an evaluation of different modeling approaches. *DP 26. Williams Project on the economics of higher education*, 1994.

[5] JK Alhassan and SA Lawal. Using data mining technique for scholarship disbursement. *World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering*, 2(7), 2015.

[6] Christopher M Antons and Elliot N Maltz. Expanding the role of institutional research at small private universities: A case study in enrollment management using data mining. *New directions for institutional research*, 2006(131):69–81, 2006.

[7] Lovenoor Aulck, Rohan Aras, Lysia Li, Coulter L'Heureux, Peter Lu, and Jevin West. Stem-ming the Tide: Predicting STEM Attrition Using Student Transcript Data. In *2017 KDD Workshop on ML4ED: Machine Learning in Education*, 2017.

[8] Lovenoor Aulck, Joshua Malters, Casey Lee, Gianni Mancinelli, Min Sun, and Jevin West. Helping Students FIG-ure It Out: A large-scale study of freshmen interest groups (FIGs) and student success. *Research in Higher Education*, 2019.

[9] Lovenoor Aulck, Joshua Malters, Casey Lee, Gianni Mancinelli, Min Sun, and Jevin West. Helping students FIG-ure it out: Using institutional records and student survey responses to examine freshmen interest groups (FIGS). *Conference of the Society for Research on Educational Effectiveness (SREE)*, 2019.

[10] Lovenoor Aulck, Dev Nambi, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Mining University Registrar Records to Predict Undergraduate Attrition and Re-Enrollment. In *Educational Data Mining Conference*, 2019.

[11] Lovenoor Aulck, Dev Nambi, and Jevin West. Optimizing scholarship allocation to improve enrollment. In *Conference on Information and Knowledge Management (CIKM)*, 2019.

[12] Lovenoor Aulck, Dev Nambi, and Jevin West. Using machine learning and genetic algorithms to optimize scholarship allocation for student yield. *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

[13] Lovenoor Aulck, Dev Nambi, and Jevin West. Using machine learning and genetic algorithms to optimize scholarship allocation for student yield. *2019 KDD Workshop on DL4ED: Deep Learning in Education*, 2019.

[14] Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Predicting Student Dropout in Higher Education. In *2016 ICML Workshop on #Data4Good: Machine Learning in Social Good Applications*, 2016.

[15] Lovenoor Aulck and Jevin West. Attrition and performance of community college transfers. *PloS One*, 12(4):e0174683, 2017.

[16] Peter C Austin. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*, 10(2):150–161, 2011.

[17] R. Baker and G. Siemens. Educational data mining and learning analytics. *The Cambridge Handbook of the Learning Sciences, Second Edition*, pages 253–272, January 2014.

[18] Ryan SJD Baker and Paul Salvador Inventado. Educational data mining and learning analytics. In *Learning Analytics*, pages 61–75. Springer, 2014.

[19] Ryan SJD Baker and Kalina Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17, 2009.

[20] Andre Baksh and Jeff E Hoyt. The effect of academic scholarships on college attendance. *College and University*, 76(4):3, 2001.

[21] Betsy O Barefoot. The first-year experience. *About Campus*, 4(6):12–18, 2000.

[22] Betsy Overman Barefoot. Helping first-year college students climb the academic ladder: Report of a national survey of freshman seminar programming in american higher education. 1993.

[23] Betsy Overman Barefoot, Carrie L. Warnock, Michael P. Dickinson, Sharon E. Richardson, and Melissa R. Roberts. Exploring the evidence: Reporting outcomes of freshman seminars. Technical Report 2, National Resource Center for the Freshman Year Experience & Students in Transition, 1998.

[24] Jaroslav Bayer, Hana Bydzovská, Jan Géryk, Tomás Obsivac, and Lubomir Popelinsky. Predicting drop-out from social behaviour of students. In *Proceedings of the 5th International Conference on Educational Data Mining*, 2012.

[25] Matthew S Berry. *The effectiveness of extended orientation first year seminars: a systematic review and meta-analysis.* PhD thesis, 2014.

[26] Marie Bienkowski, Mingyu Feng, and Barbara Means. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. US Department of Education Office of Educational Technology, 2012.

[27] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[28] Andreas Blom and Erik Canton. *Can student loans improve accessibility to higher education and student performance? An impact study of the case of SOFES, Mexico.* The World Bank, 2004.

[29] William Emerson Brock. *An American Imperative: Higher Expectations for Higher Education: A Report.* Johnson Foundation, 1993.

[30] Brad Brown, Jacques Bugin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: the next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, 2011.

[31] Donald J Bruce and Celeste K Carruthers. Jackpot? the impact of lottery scholarships on enrollment in tennessee. *Journal of Urban Economics*, 81:30–44, 2014.

[32] Alberto F Cabrera, Amaury Nora, and Maria B Castaneda. The role of finances in the persistence process: A structural model. *Research in Higher Education*, 33(5):571–593, 1992.

[33] Alberto F Cabrera, Amaury Nora, and Maria B Castaneda. College persistence: Structural equations modeling test of an integrated model of student retention. *The journal of higher education*, 64(2):123–139, 1993.

[34] Alberto F Cabrera, Jacob O Stampen, and W Lee Hansen. Exploring the effects of ability to pay on persistence in college. *The Review of Higher Education*, 13(3):303–336, 1990.

[35] Amy L Caison. Analysis of institutionally specific retention research: A comparison between survey and institutional database methods. *Research in Higher Education*, 48(4):435–451, 2007.

[36] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72, 2008.

[37] John P Campbell, Peter B DeBlois, and Diana G Oblinger. Academic analytics: A new tool for a new era. *EDUCAUSE review*, 42(4):40, 2007.

[38] Anthony P Carnevale, Nicole Smith, and Jeff Strohl. Help wanted: projections of jobs and education requirements through 2018. 2010.

[39] Anthony P Carnevale, Nicole Smith, and Jeff Strohl. Recovery: Job growth and education requirements through 2020. 2013.

[40] Lin Chang. Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research*, 2006(131):53–68, 2006.

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

[42] MH Clark and Nicole L Cundiff. Assessing the effectiveness of a college freshman seminar using propensity score adjustments. *Research in Higher Education*, 52(6):616–639, 2011.

[43] Sarah R Cohodes and Joshua S Goodman. Merit aid, college quality, and college completion: Massachusetts' adams scholarship as an in-kind subsidy. *American Economic Journal: Applied Economics*, 6(4):251–85, 2014.

[44] Michael D Coomes. The historical roots of enrollment management. *New directions for student services*, 2000(89):5–18, 2000.

[45] Christopher Cornwell, David B Mustard, and Deepa J Sridhar. The enrollment effects of merit-based financial aid: Evidence from georgia's hope program. *Journal of Labor Economics*, 24(4):761–786, 2006.

[46] Christopher M Cornwell, Kyung Hee Lee, and David B Mustard. Student responses to merit scholarship retention rules. *Journal of Human Resources*, 40(4):895–917, 2005.

[47] Jennifer Crissman. Clustered and nonclustered first-year seminars: New students' first-semester experiences. *Journal of the First-Year Experience & Students in Transition*, 13(1):69–88, 2001.

[48] Jennifer L Crissman. The impact of clustering first year seminars with english composition courses on new students' retention rates. *Journal of College Student Retention: Research, Theory & Practice*, 3(2):137–152, 2001.

[49] Bradley R Curs and Casandra E Harper. Financial aid and first-year collegiate gpa: A regression discontinuity approach. *The Review of Higher Education*, 35(4):627–649, 2012.

[50] Bradley R Curs and Larry D Singell Jr. Aim high or go low? pricing strategies and enrollment effects when the net price elasticity varies with need and ability. *The Journal of Higher Education*, 81(4):515–543, 2010.

[51] Joseph B Cuseo. Freshman orientation seminar at community colleges: A research-based rationale for its value, content, and delivery. 1997.

[52] Joseph B Cuseo. The empirical case for the first-year seminar: Promoting positive student outcomes and campus-wide benefits. In *The first-year seminar: Research-based recommendations for course design, delivery, and assessment*. Kendall/Hunt Dubuque, IA, 2010.

[53] Dawn Darlaston-Jones, Lynne Cohen, Suena Haunold, Lisbeth Pike, Alison Young, and Neil Drew. The retention and persistence support (raps) project: A transition initiative. 2003.

[54] Brandon De la Cuesta and Kosuke Imai. Misunderstandings about the regression discontinuity design in the study of close elections. *Annual Review of Political Science*, 19:375–396, 2016.

[55] Gerben W Dekker, Mykola Pechenizkiy, and Jan M Vleeshouwers. Predicting students drop out: A case study. *International Working Group on Educational Data Mining*, 2009.

[56] Dursun Delen. Predicting student attrition with data mining methods. *Journal of College Student Retention: Research, Theory & Practice*, 13(1):17–35, 2011.

[57] Cynthia Demetriou and Amy Schmitz-Sciborski. Integration, motivation, strengths and optimism: Retention theories past, present and future. In *Proceedings of the 7th National Symposium on Student Retention, Charleston, SC*, pages 300–312, 2011.

[58] Stephen L DesJardins. An analytic strategy to assist institutional recruitment and marketing efforts. *Research in Higher education*, 43(5):531–553, 2002.

[59] 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.

[60] Vasant Dhar. Data science and prediction. *Communications of the ACM*, 56(12):64–73, 2013.

[61] John Aubrey Douglass. Higher education budgets and the global recession: Tracking varied national responses and their consequences. research & occasional paper series: Cshe. 4.10. *Center for Studies in Higher Education*, 2010.

[62] Raymond W Drake. Review of the literature for freshman orientation practices in the US. 1966.

[63] Susan Dynarski. Hope for whom? financial aid for the middle class and its impact on college attendance. Technical report, National bureau of economic research, 2000.

[64] Ronald G Ehrenberg and Daniel R Sherman. Optimal financial aid policies for a selective university. *Journal of Human Resources*, pages 202–230, 1984.

[65] Ronald G Ehrenberg, Liang Zhang, and Jared Levin. Crafting a class: The trade off between merit scholarships and enrolling lower-income students. Technical report, National Bureau of Economic Research, 2005.

[66] Paul Fidler. Relationship of freshman orientation seminars to sophomore return rates. *Journal of the First-Year Experience & Students in Transition*, 3(1):7–38, 1991.

[67] David S Fike and Renea Fike. Predictors of first-year student retention in the community college. *Community college review*, 36(2):68–88, 2008.

[68] Félix-Antoine Fortin, François-Michel De Rainville, Marc-André Gardner, Marc Parizeau, and Christian Gagné. Deap: Evolutionary algorithms made easy. *Journal of Machine Learning Research*, 13(Jul):2171–2175, 2012.

[69] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[70] Josephine A Gasiewski, M Kevin Eagan, Gina A Garcia, Sylvia Hurtado, and Mitchell J Chang. From gatekeeping to engagement: A multicontextual, mixed method study of student academic engagement in introductory stem courses. *Research in Higher Education*, 53(2):229–261, 2012.

[71] Cullen F Goenner and Kenton Pauls. A predictive model of inquiry to enrollment. *Research in Higher education*, 47(8):935–956, 2006.

[72] Julie M Byers González and Stephen L DesJardins. Artificial neural networks: A new approach to predicting application behavior. *Research in Higher Education*, 43(2):235–258, 2002.

[73] Joshua Goodman. Who merits financial aid?: Massachusetts' adams scholarship. *Journal of public Economics*, 92(10-11):2121–2131, 2008.

[74] Kathleen Goodman and Ernest T Pascarella. First-year seminars increase persistence and retention: A summary of the evidence from how college affects students. *Peer Review*, 8(3):26, 2006.

[75] Angela M Griffin and Jonathan Romm. Exploring the evidence: Reporting research on first-year seminars. Technical Report 4, National Resource Center for The First-Year Experience & Students in Transition, 2008.

[76] Heidi E Grunwald and Matthew J Mayhew. Using propensity scores for estimating causal effects: A study in the development of moral reasoning. *Research in Higher Education*, 49(8):758–775, 2008.

[77] Sherif Halawa, Daniel Greene, and John Mitchell. Dropout prediction in MOOCs using learner activity features. *Experiences and best practices in and around MOOCs*, 7, 2014.

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

[79] Nicolaus Henke, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, Bill Wiseman, and Guru Sethupathy. The age of analytics: Competing in a data-driven world. *McKinsey Global Institute*, page 4, 2016.

[80] John Hood. The new austerity: University budgets in the 1990s. *Academic Questions*, 9(2):82–88, 1996.

[81] David S Hopkins. *Planning models for colleges and universities*. Stanford University Press, 1981.

[82] Laura Horn and C Dennis Carroll. Stopouts or stayouts. Undergraduates who leave college in their first year. Technical report, National Center for Education Statistics, 1998.

[83] Don Hossler. The role of financial aid in enrollment management. *New directions for student services*, 2000(89):77–90, 2000.

[84] Don Hossler. Managing student retention: Is the glass half full, half empty, or simply empty? *College and University*, 81(2):11–14, 2006.

[85] Don Hossler. Enrollment management & the enrollment industry. *College and University*, 85(2):2, 2009.

[86] Harold A Hovey. State spending for higher education in the next decade: The battle to sustain current support. 1999.

[87] William E Hudson Sr. Can an early alert excessive absenteeism warning system be effective in retaining freshman students? *Journal of College Student Retention: Research, Theory & Practice*, 7(3):217–226, 2005.

[88] Robin Jacob, Pei Zhu, Marie-Andrée Somers, and Howard Bloom. A practical guide to regression discontinuity. *MDRC*, 2012.

[89] Eric Jamelske. Measuring the impact of a university first-year experience program on student GPA and retention. *Higher Education*, 57(3):373–391, 2009.

[90] Sandeep M Jayaprakash, Erik W Moody, Eitel JM Lauría, James R Regan, and Joshua D Baron. Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, 1(1):6–47, 2014.

[91] Nate Johnson. The institutional costs of student attrition. *Delta Cost Project at American Institutes for Research*, 2012.

[92] Thomas J Kane. A quasi-experimental estimate of the impact of financial aid on college-going. Technical report, National Bureau of Economic Research, 2003.

[93] Grace Kena, William Hussar, Joel McFarland, Cristobal de Brey, Lauren Musu-Gillette, Xiaolei Wang, Jijun Zhang, Amy Rathbun, Sidney Wilkinson-Flicker, Melissa Diliberti, et al. The condition of education 2016. (NCES 2016-144). Technical report, National Center for Education Statistics, 2016.

[94] Shirley Strum Kenny, Bruce Alberts, Wayne C. Booth, Milton Glaser, Charles E. Glassick, Stanley O. Ikenberry, Kathleen Hall Jamieson, Robert M. O'Neil, Carolynn Reid-Wallace, Chang-Lin Tien, and Chen Ning Yang. Reinventing undergraduate education: A blueprint for America's research universities. Technical report, Boyer Commission on Educating Undergraduates in the Research University, 1998.

[95] Jennifer Keup and Betsy O. Barefoot. Learning how to be a successful student: Exploring the impact of first-year seminars on student outcomes. *Journal of The First-Year Experience & Students in Transition*, 17(1):11–47, 2005.

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

[97] Gary King and Richard Nielsen. Why propensity scores should not be used for matching. 2016.

[98] J Klatt and R Ray. Student academic outcomes after completing a first-year seminar. *NACTA Journal*, 58(4), 2014.

[99] Zlatko J Kovačić. Early prediction of student success: mining students enrolment data. In *Proceedings of Informing Science & IT Education Conference (InSITE)*, pages 647–665. Citeseer, 2010.

[100] Robert B Kozma. Technology and classroom practices: An international study. *Journal of research on technology in education*, 36(1):1–14, 2003.

[101] David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

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

[103] Larry L Leslie and Paul T Brinkman. *The Economic Value of Higher Education. American Council on Education/Macmillan Series on Higher Education.* ERIC, 1988.

[104] JJ Lin, PK Imbrie, and Kenneth J Reid. Student retention modelling: An evaluation of different methods and their impact on prediction results. *Research in Engineering Education Sysmposium*, pages 1–6, 2009.

[105] Gordon S. Linoff and Michael J. A. Berry. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Wiley, 2011.

[106] Yuetian Luo and Zachary A Pardos. Diagnosing university student subject proficiency and predicting degree completion in vector space. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[107] Ioanna Lykourentzou, Ioannis Giannoukos, Vassilis Nikolopoulos, George Mpardis, and Vassili Loumos. Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, 53(3):950–965, 2009.

[108] Rubén Manrique, Bernardo Pereira Nunes, Olga Marino, Marco Antonio Casanova, and Terhi Nurmikko-Fuller. An analysis of student representation, representative features and classification algorithms to predict degree dropout. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 401–410. ACM, 2019.

[109] Joel McFarland, Bill Hussar, Cristobal de Brey, Tom Snyder, Xiaolei Wang, Sidney Wilkinson-Flicker, Semhar Gebrekristos, Jijun Zhang, Amy Rathbun, Amy Barmer, et al. The condition of education 2017. (NCES 2017-144). Technical report, National Center for Education Statistics, 2017.

[110] Michael K McLendon, David A Tandberg, and Nicholas W Hillman. Financing college opportunity: Factors influencing state spending on student financial aid and campus appropriations, 1990 through 2010. *The ANNALS of the American Academy of Political and Social Science*, 655(1):143–162, 2014.

[111] Michael S McPherson and Morton Owen Schapiro. Does student aid affect college enrollment? new evidence on a persistent controversy. *The American Economic Review*, 81(1):309–318, 1991.

[112] Charles Miller, Nicholas Donofrio, James J. Duderstadt, Gerri Elliott, Jonathan N. Grayer, Kati Haycock, James B. Hunt Jr., Arturo Madrid, Robert Mendenhall, Charlene R. Nunley, Catherine B. Reynolds, Arthur J. Rothkopf, Richard Stephens, Louis W. Sullivan, Sara Martinez Tucker, Richard Vedder, Charles M. Vest, and Robert M. Zemsky. A test of leadership: Charting the future of us higher education. Technical report, US Department of Education, 2006.

[113] M Millikin. Promoting student success: Evaluation of a freshman orientation course. *Innovation and Empowerment: SNU-Tulsa Research Journal*, 3(2):1–14, 2011.

[114] Melanie Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

[115] Shannon Mok and Joshua Shakin. Distribution of federal support for students pursuing higher education in 2016. *Congressional Budget Office*, 2018.

[116] James Monks. The impact of merit-based financial aid on college enrollment: A field experiment. *Economics of Education Review*, 28(1):99–106, 2009.

[117] Claude Montmarquette, Kathy Cannings, and Sophie Mahseredjian. How do young people choose college majors? *Economics of Education Review*, 21(6):543–556, 2002.

[118] Gordon E Moore et al. Cramming more components onto integrated circuits, 1965.

[119] Peter Moran. Reacting to crises: The risk-averse nature of contemporary american public education. *Policy futures in education*, 13(5):621–638, 2015.

[120] Kenneth P Mortimer. Involvement in learning: Realizing the potential of American higher education. Technical report, National Institute of Education, 1984.

[121] Laurence G Moseley and Donna M Mead. Predicting who will drop out of nursing courses: a machine learning exercise. *Nurse education today*, 28(4):469–475, 2008.

[122] John A Muffo and Gerald W McLaughlin. A primer on institutional research. Association for Institutional Research, 1987.

[123] Marcell Nagy and Roland Molontay. Predicting dropout in higher education based on secondary school performance. In *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, pages 000389–000394. IEEE, 2018.

[124] Ashutosh Nandeshwar and Subodh Chaudhari. Enrollment prediction models using data mining. *Retrieved January*, 10:2010, 2009.

[125] David Niemi and Elena Gitin. Using big data to predict student dropouts: Technology affordances for research. In *Proceedings of the International Association for Development of the Information Society (IADIS) International Conference on Cognition and Exploratory Learning in Digital Age*, 2012.

[126] Ryan D Padgett, Jennifer R Keup, and Ernest T Pascarella. The impact of first-year seminars on college students life-long learning orientations. *Journal of Student Affairs Research and Practice*, 50(2):133–151, 2013.

[127] Timothy J Pantages and Carol F Creedon. Studies of college attrition: 1950–1975. *Review of educational research*, 48(1):49–101, 1978.

[128] Ernest T Pascarella and Patrick T Terenzini. *How college affects students*, volume 2. Jossey-Bass San Francisco, CA, 2005.

[129] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[130] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568. ACM, 2008.

[131] 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.

[132] Sameano F Porchea, Jeff Allen, Steve Robbins, and Richard P Phelps. Predictors of long-term enrollment and degree outcomes for community college students: Integrating academic, psychosocial, socio-demographic, and situational factors. *The Journal of higher education*, 81(6):750–778, 2010.

[133] Stephen R Porter and Randy L Swing. Understanding how first-year seminars affect persistence. *Research in Higher Education*, 47(1):89–109, 2006.

[134] John R Purdie and Vicki J Rosser. Examining the academic performance and retention of first-year students in living-learning communities and first-year experience courses. *College Student Affairs Journal*, 29(2):95, 2011.

[135] Neal Raisman. The cost of college attrition at four-year colleges & universities. policy perspectives. Technical report, Educational Policy Institute, 2013.

[136] Sudha Ram, Yun Wang, Faiz Currim, and Sabah Currim. Using big data for predicting freshmen retention. 2015.

[137] Cristóbal Romero and Sebastián Ventura. Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146, 2007.

[138] Cristóbal Romero and Sebastián Ventura. Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618, 2010.

[139] Cristobal Romero and Sebastian Ventura. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27, 2013.

[140] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[141] Sudeepa Roy, Cynthia Rudin, Alexander Volfovsky, and Tianyu Wang. Flame: A fast large-scale almost matching exactly approach to causal inference. *arXiv preprint arXiv:1707.06315*, 2017.

[142] Donald B Rubin and Neal Thomas. Matching using estimated propensity scores: relating theory to practice. *Biometrics*, pages 249–264, 1996.

[143] Rachel B Rubin. The pell and the poor: A regression-discontinuity analysis of on-time college enrollment. *Research in Higher Education*, 52(7):675–692, 2011.

[144] S Sadati and Nicolas Ali Libre. Development of an early alert system to predict students at risk of failing based on their early course activities. 2017.

[145] Arun P Sanjeev and Jan M Zytkow. Discovering enrollment knowledge in university databases. In *KDD*, pages 246–251, 1995.

[146] Z Sarafraz, H Sarafraz, M Sayeh, and J Nicklow. Student yield maximization using genetic algorithm on a predictive enrollment neural network model. *Procedia Computer Science*, 61:341–348, 2015.

[147] Marlene Scardamalia. Big change questions will educational institutions, within their present structures, be able to adapt sufficiently to meet the needs of the information age? *Journal of Educational Change*, 2(2):171–176, 2001.

[148] Mark Schneider. Finishing the first lap: The cost of first year student attrition in America's four year colleges and universities. *American Institutes for Research*, 2010.

[149] Carolyn A Schnell and Curt D Doetkott. First year seminars produce long-term impact. *Journal of College Student Retention: Research, Theory & Practice*, 4(4):377–391, 2003.

[150] Carolyn A Schnell, Karen Louis, and Curt D Doetkott. The first-year seminar as a means of improving college graduation rates. *Journal of The First-Year Experience & Students in Transition*, 15(1):53–76, 2003.

[151] Jasjeet S Sekhon and Rocío Titiunik. On interpreting the regression discontinuity design as a local experiment. In *Regression discontinuity designs: Theory and applications*, pages 1–28. Emerald Publishing Limited, 2017.

[152] Joseph J Seneca and Michael K Taussig. The effects of tuition and financial aid on the enrollment decision at a state university. *Research in Higher Education*, 26(4):337–362, 1987.

[153] Xanthe Shacklock. *From bricks to clicks: the potential of data and analytics in higher education*. Higher Education Commission London, 2016.

[154] Raj Man Shrestha, Mehmet A Orgun, and Peter Busch. Offer acceptance prediction of academic placement. *Neural Computing and Applications*, 27(8):2351–2368, 2016.

[155] George Siemens. Learning analytics: envisioning a research discipline and a domain of practice. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, pages 4–8. ACM, 2012.

[156] George Siemens and Phil Long. Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, 46(5):30, 2011.

[157] Jill M Simons. *A national study of student early alert models at four-year institutions of higher education*. Arkansas State University, 2011.

[158] Larry D Singell Jr. Come and stay a while: does financial aid effect retention conditioned on enrollment at a large public university? *Economics of Education review*, 23(5):459–471, 2004.

[159] Fadzilah Siraj and Mansour Ali Abdoulha. Uncovering hidden information within university's student enrollment data using data mining. In *Modelling & Simulation, 2009. AMS'09. Third Asia International Conference on*, pages 413–418. IEEE, 2009.

[160] David L Sjoquist and John V Winters. State merit-based financial aid programs and college attainment. *Journal of Regional Science*, 55(3):364–390, 2015.

[161] Christopher Skovron and Rocıo Titiunik. A practical guide to regression discontinuity designs in political science. 2015.

[162] William Spady. Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1):64–85, 1970.

[163] William Spady. Dropouts from higher education: Toward an empirical model. *Interchange*, 2(3):38–62, 1971.

[164] Randy Spaulding and Steven Olswang. Maximizing enrollment yield through financial aid packaging policies. *Journal of Student Financial Aid*, 35(1):3, 2005.

[165] Rodney E Stanley and P Edward French. Evaluating increased enrollment levels in institutions of higher education: A look at merit-based scholarship programs. *Public Administration Quarterly*, pages 4–36, 2009.

[166] Martha LA Stassen. Student outcomes: The impact of varying living-learning community models. *Research in Higher Education*, 44(5):581–613, 2003.

[167] Gerry Strumpf and Pat Hunt. The effects of an orientation course on the retention and academic standing of entering freshmen, controlling for the volunteer effect. *Journal of The First-Year Experience & Students in Transition*, 5(1):7–14, 1993.

[168] Claire F Sullivan and DH Wulff. Freshman interest groups at the University of Washington: Building community for freshmen at a large university. *Washington Center News*, 4:1–8, 1990.

[169] John Summerskill. Dropouts from college. In *The American College*. Wiley, New York, 1965.

[170] Watson Scott Swail, Kenneth E. Redd, and Laura W. Perna. Retaining minority students in higher education: A framework for success. Technical Report 2, ASHE-ERIC, 2003.

[171] Dawn Geronimo Terkla. Does financial aid enhance undergraduate persistence? *Journal of Student Financial Aid*, 15(3), 1985.

[172] Dech Thammasiri, Dursun Delen, Phayung Meesad, and Nihat Kasap. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, 41(2):321–330, 2014.

[173] Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309, 1960.

[174] Vincent Tinto. Defining dropout: A matter of perspective. *New Directions for Institutional Research*, 1982(36):3–15, 1982.

[175] Vincent Tinto. *Leaving college: Rethinking the causes and cures of student attrition.* University of Chicago Press, 1987.

[176] Vincent Tinto. Taking retention seriously: Rethinking the first year of college. *NACADA journal*, 19(2):5–9, 1999.

[177] Vincent Tinto. Learning better together: The impact of learning communities on student success. *Higher Education Monograph Series*, 1(8):1–8, 2003.

[178] Vincent. Tinto and Anne Goodsell. *A Longitudinal Study of Freshman Interest Groups at the University of Washington.* Distributed by ERIC Clearinghouse, 1993.

[179] Vincent Tinto and Anne Goodsell. Freshman interest groups and the first-year experience: Constructing student communities in a large university. *Journal of The First-Year Experience & Students in Transition*, 6(1):7–28, 1994.

[180] Kenneth Tokuno. Long-term and recent student outcomes of the freshman interest group program. *Journal of the First-Year Experience & Students in Transition*, 5(2):7–28, 1993.

[181] Kenneth Tokuno and Frederick Campbell. The freshman interest group program at the University of Washington: Effects on retention and scholarship. *Journal of The First-Year Experience & Students in Transition*, 4(1):7–22, 1992.

[182] Dale Trusheim and Carol Rylee. Predictive modeling: linking enrollment and budgeting. *Planning for Higher Education*, 40(1):12, 2011.

[183] M. Lee Upcraft and John N. Gardner. *The freshmen year experience: helping students survive and succeed in college.* Jossey-Bass, 1989.

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

[185] E Velez. The condition of education 2015 (NCES 2015-144). Technical report, National Center for Education, 2015.

[186] Steven Walczak. Neural network models for a resource allocation problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 28(2):276–284, 1998.

[187] Steven Walczak and Terry Sincich. A comparative analysis of regression and neural networks for university admissions. *Information Sciences*, 119(1-2):1–20, 1999.

[188] Darrell M West. Big data for education: Data mining, data analytics, and web dashboards. 2012.

[189] A Michael Williford, Laura Cross Chapman, and Tammy Kahrig. The university experience course: A longitudinal study of student performance, retention, and graduation. *Journal of College Student Retention: Research, Theory & Practice*, 2(4):327–340, 2001.

[190] Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-driven education workshop*, volume 11, page 14, 2013.