STRUCTURAL NEURAL NETWORKS MEET PIECEWISE EXPONENTIAL MODELS ON CUSTOMER ACQUISITION AND RETENTION

by

Chuan Cai

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Financial Services Analytics

Winter 2024

STRUCTURAL NEURAL NETWORKS MEET PIECEWISE EXPONENTIAL MODELS ON CUSTOMER ACQUISITION AND RETENTION

by

Chuan Cai

Approved: ____

Bintong Chen, Ph.D.
Chair of the Institute of Financial Services Analytics

Approved: ____

Oliver Yao, Ph.D.
Dean of the College of Business and Economics

Approved: ____

Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ______________________________

Adam Fleischhacker, Ph.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ______________________________

Bintong Chen, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ______________________________

Keith Decker, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed: ______________________________

Wei Qian, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

I am profoundly indebted to my advisor, Dr. Adam Fleischhacker. His unwavering patience, support, and encouragement have been instrumental in the fruition of this work. Learning from him about integrating domain knowledge with statistical models to address real-world challenges has been an enlightening journey. I consider it a distinct honor to be his mentee. His sagacity and optimism guide not just my research endeavors but also illuminate many aspects of my life.

My heartfelt gratitude extends to my dissertation committee: Dr. Bintong Chen for his overarching perspective on the projects; Dr. Wei Qian for his invaluable input on modeling techniques; and Dr. Keith Decker for his incisive comments on the proposal.

To Richard Reeves, my supervisor at the Office of Institutional Research and Effectiveness, I extend my sincere appreciation. Your unwavering support in my pursuit of this degree has been pivotal. Our discussions on student enrollment management sparked the inspiration for my research projects, and your expertise has immensely broadened my understanding in this domain.

Last, but certainly not least, my deepest gratitude is reserved for my family. To my wife, Shuang Ru, your unwavering love and support provided the foundation upon which this work stands. I am eternally grateful for the sacrifices you've made, allowing me to dedicate myself to this pursuit. To our cherished daughter, Alice Yiru Cai, I pledge more moments of joy, laughter, and cherished memories together. To my

parents and in-laws, your dedication and care for Alice during this period have been a bedrock of support. I am profoundly grateful.

TABLE OF CONTENTS

Appendix

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This dissertation examines the dynamics of customer acquisition and retention within the context of higher education with students being recognized as customers. Given the substantial reliance of higher education institutions on tuition fees for operating revenue, student recruitment and retention emerge as pivotal aspects of enrollment management. Furthermore, retention and timely graduation are used to measure institutional reputation and accountability. This study pursues three main objectives: understanding applicant deposit decisions during the admission process, predicting matriculated students' dropout risks, and exploring the impact of student loan debt on timely graduation.

The first segment analyzes the determinants of deposit decisions of out-of-state students admitted to the University of Delaware across three academic years. Utilizing three Bayesian hierarchical piecewise exponential models, we deduce that factors like gender and recruitment events exhibit time-varying effects, while others like financial aid remain stable within an academic year but vary across years. The baseline desire to deposit intensifies as deadlines near, though this trajectory shifts annually. Insights derived inform the Admissions Office's marketing and recruitment tactics.

The second segment introduces a hybrid model, merging a structural neural network with a piecewise exponential model, to predict college attrition. Benchmarking against two alternative models, the hybrid model demonstrates superior or comparable predictive prowess for the University of Delaware across three springs. Categorizing predictors into academic, economic, and socio-demographic facets reveals academic indicators as key discriminants between students who drop out and those retained, especially from freshman to junior years. Emphasis on academic assessments in intervention strategies is thus recommended.

The third segment evaluates the impact of student loan debt on six-year graduation rates by department, over a span of five years. Leveraging five Bayesian hierarchical models, the findings illustrate a pronounced department-wise loan debt effect on first-year students, which attenuates as they advance academically. Tailored financial aid policies, considering academic departments, are posited to amplify the efficient utilization of institutional financial resources. For universities mulling over department-specific financial aid policies, initiation with randomized trials for first-year students is advised.

In summary, this dissertation introduces innovative strategies for strategic enrollment management, encompassing admission, retention, and graduation considerations. Particular attention is given to the dynamic nature of applicant deposit decisions, the development of predictive models for student attrition, and the department-specific effects of student loan debt on graduation rates.

Chapter 1

INTRODUCTION

Customer acquisition and retention are critical components of customer relationship management (CRM), prevalent across various industries including hotel management, banking, insurance, healthcare, higher education, and agriculture (Furman et al., 2021; Milovic, 2012). Effective customer acquisition strategies aid in crafting marketing campaigns that recognize, attract, and recruit new customers early on. Conversely, robust customer retention strategies nurture both short-term and long-term relationships with existing clientele, augmenting their loyalty and satisfaction. Additionally, analyzing customer retention provides insights into dissatisfied customers, enabling businesses to discern unmet needs and enhance customer experiences.

Within the higher education sector, the concept of CRM has garnered attention over the years, with students increasingly recognized as valuable customers (Mashenene et al., 2019; Ogunnaike et al., 2014; Tapp et al., 2004). As an institutional researcher in a higher education institution, I focus this study of customer acquisition and retention on the higher education landscape. In this context, customer acquisition and retention take on the roles of student recruitment and retention, well-established concepts in the domain of student enrollment management (SEM) (Nair et al., 2007; Virgiyanti et al., 2010). This dissertation comprises three pivotal subjects within SEM: understanding applicants' deposit decisions, predicting enrolled students' dropout risk, and discerning the influence of loan debt on graduation. Each subsequent chapter is dedicated to one of these topics, ensuring that the examination of these multifaceted issues is thorough and comprehensive.

Chapter 2 of this dissertation responds to the pressing need within higher education institutions, particularly in their Admissions Offices, to gain a deeper understanding of deposit decisions made by applicants. Factors affecting these decisions, such as the offering of financial aid, demographic characteristics, and admission events, are examined through the lens of Bayesian hierarchical piecewise exponential models. These models are tailored to capture the longitudinal decision-making process of students between February and May, divided into eight distinct time periods. They facilitate the estimation of both the baseline desire to make deposits and the time-varying effects of various influencing factors. The introduction of a Bayesian hierarchical framework is a strategic move to strike the right balance between underfitting and overfitting, particularly concerning the time-varying effects. This investigation primarily focuses on the effects of various factors on out-of-state students' deposit decisions who applied to the University of Delaware and were admitted for the Fall semesters of 2020, 2021, and 2022.

Chapter 3 seeks to contribute to the design of intervention programs aimed at addressing the issue of college student attrition. Our approach involves modeling students' academic journeys using piecewise exponential models, incorporating academic, economic, and socio-demographic factors. To achieve a balance between predictivity and interpretability, we introduce structural neural networks to model the hazard function within the piecewise exponential models. This hybrid model serves not only to predict students at high risk of dropout but also to shed light on whether this risk primarily emanates from academic, economic, or socio-demographic factors.

Comparative analyses with other models, including piecewise exponential models and a hybrid model combining a fully-connected neural network with piecewise exponential models, demonstrate the superiority or comparability of the hybrid structural neural network model in predicting dropout among University of Delaware students in the Spring semesters of 2020, 2021, and 2022. Moreover, our findings consistently highlight the paramount role of academic factors in discerning dropout risks, regardless of the stage of students' academic journeys.

In Chapter 4, our attention shifts to exploring the impact of student loan debt on graduation, with the aim of informing potential modifications to financial aid policies to enhance six-year graduation rates. Of particular interest is the examination of whether the effects of loan debt remain consistent across students majoring in different departments. We employ Bayesian hierarchical logit models to estimate department-level effects, considering that some departments have relatively small enrollments. The introduction of a Bayesian hierarchical structure mitigates overfitting concerns, with college-level effects serving as higher-level factors. Drawing from data spanning student cohorts from Fall 2009 to Fall 2011 at the University of Delaware, we uncover variations in the effects of loan debt among departments, particularly among first-year students. Importantly, we note that these differences tend to diminish as students progress through their academic careers. We posit that universities should contemplate the exploration of department-specific financial aid policies, with randomized trials among first-year students as an auspicious starting point.

In summary, this dissertation delves into the intricate landscape of strategic enrollment management, spanning the domains of admission, retention, and graduation. Chapter by chapter, we probe into the dynamics of applicants' deposit decisions, the prediction of students' dropout risks, and the influence of loan debt on graduation rates, offering insights and recommendations that have the potential to shape more effective practices within the realm of higher education.

Chapter 2

WHEN THEY WILL PAY: UNDERSTAND DEPOSIT DECISIONS IN COLLEGE ADMISSION

2.1 Introduction

Recruitment of new students is a major challenge for many higher education institutions, as they must balance revenue goals, academic standards, and diversity goals (Adams-Johnson et al. 2019; Maldonado, Armelini, and Guevara 2017). The recruitment process includes several steps, such as answering inquiries from prospective students, reviewing applications, making admission decisions, encouraging deposit payments, and assisting with matriculation (Litten et al. 1983). Although matriculation is the ultimate goal, deposit payments are a critical focus for the Admissions Office, as they are a strong indicator of a student's likelihood to enroll. Other offices/departments in an institution also have a vested interest in deposit situation. For example, the Budget Office would like to use deposits paid to estimate the tuition revenue from new students. If the deposits paid differs greatly from the admission targets, the Budget Office must update budgetary models. Similarly, the Department of Residence Life and Housing needs to work on the potential shortage of residence hall space in the spring, if too many deposits are paid. As a result, the Admissions Office tracks deposits usually by week in the spring and desires to understand admitted students' decision to pay deposits, not only whether they will pay, but also when they will pay, so they can adjust marketing and recruitment strategies promptly (DesJardins 2002; Goenner and Pauls 2006).

Previous studies have greatly contributed to our understanding of student enrollment decisions and have informed the development of marketing and recruitment strategies. DesJardins (2002) utilized predictive models to segment admitted students and recommended focusing on those with enrollment probabilities close to 0.5. Maldonado et al. (2017) developed nested logit models to predict the enrollment probabilities of admitted students, allowing decision makers to allocate resources for admission activities. Goenner and Pauls (2006) used predictive models to determine the enrollment probabilities of prospective students, helping the University of North Dakota allocate recruitment efforts by geographic area. Braunstein et al. (1999) studied the impact of financial factors on the enrollment decisions of admitted students and provided insight into how various types of financial aid affected students from different socio-economic backgrounds. Johnson (2019) investigated the factors that influenced the enrollment decisions of out-of-state students, identifying multiple potential destinations and providing insights into why students accepted or declined an offer of admission.

Our study focuses on understanding the factors affecting when admitted students will pay their deposits, as opposed to simply whether they will enroll by using the cross-sectional methods used in previous studies. To accomplish this, we utilize event history analysis, which is a commonly used tool for modeling students' journey from enrollment to graduation (Chen and Hossler 2017; Gross, Torres, and Zerquera 2013; Zhan, Xiang, and Elliott III 2018). However, we use piecewise exponential models for the event history analysis instead of proportional hazard models for two reasons. First, we desire to estimate students' baseline deposit behavior regardless of students' characteristics. Second, we assume that it is possible that the factors related to deposit decisions have time-varying effects. Additionally, we employ a Bayesian hierarchical framework to balance overfitting and underfitting of the time-varying effects (McElreath 2020). The study uses three years of admission data from 2020 to 2022 provided by the Admissions Office at the University of Delaware (UD). The input variables for the Bayesian hierarchical piecewise exponential models were gathered through a review of previous studies (Paulsen 1990) and suggestions from the Admissions Office. These variables can be grouped into economic factors (e.g. offered financial aid), sociological factors (e.g. gender), and psychological factors (e.g. delay in reviewing admission decisions).

This chapter aims to address the following two research questions in order to understand the factors that impact admitted students' deposit decisions and support the Admissions Office's recruitment efforts:

1. Do the factors influencing deposit decisions exhibit varying effects over the course of a year?

2. Do effects change from one year to the next?

2.2 Theoretical Background and Practical Application

2.2.1 Admission Funnel

From the perspective of the institution, the admission process consists of six stages: prospects, inquirers, applicants, admitted students, depositors, and matriculants (Litten et al. 1983). In the prospect stage, the Admissions Office identifies potential students who may be interested in attending the institution. During the inquiry stage, the Admissions Office communicates with students who have expressed interest and works to increase their interest in the institution and encourage them to apply. In the application stage, the Admissions Office processes and reviews completed applications and notifies students of any missing information. In the admit stage, the Admissions Office decides to offer admission, waitlist, or reject applications. During the deposit stage, the Admissions Office interacts with admitted students through campus tours and other admission events, and the Student Financial Services (SFS) office provides financial aid packages in an effort to encourage students to accept the offer of admission. In the enrollment stage, the Admissions Office works with other offices to support new student orientation, course registration, and on-campus residency. There are several key rates to monitor in the admission process, including conversion rate (the proportion of applicants from inquiries), selection rate (the proportion of admitted students from applicants), yield rate (the proportion of depositors from admitted students), and melt rate (the proportion of matriculants from depositors). At the University of Delaware (UD), with the melt rate typically close to 100%, the deposit stage or yield rate largely determines the number of new students enrolling each fall.

2.2.2 College Choice

From a student's perspective, the college choice process consists of three stages: college aspiration formation, search and application, and selection and attendance (Hossler and Gallagher 1987; Paulsen 1990). In the formation stage, students determine whether to pursue higher education, with factors such as family background (Stage and Hossler 1989; Carpenter and Fleishman 1987), teacher and counselor encouragement (Portes and Wilson 1976; Conklin and Dailey 1981), and academic aptitude and achievement (Tuttle 1981; Davies and Kandel 1981) influencing their decision. During the search and application stage, students compile a list of colleges to apply to, typically starting in junior year of high school and completing applications in senior year (Gilmour Jr et al. 1981). In addition to input from parents, counselors, and peers, colleges also reach out to prospective students through publications such as guidebooks and campus events (Goenner and Pauls 2006). In the final stage, selection and attendance, students make the decision of which college to attend, based on their demographic background, socio-economic status, academic preparation, and institution characteristics such as cost, financial aid, academic programs, academic reputation, and location (DesJardins 2002; Goenner and Pauls 2006).

Many studies have used cross-sectional methods to examine students' enrollment decisions in the selection and attendance stage. DesJardins (2002) employed a logistic regression model to predict the enrollment probabilities of students admitted to a public institution in the Midwest in 1999 and 2001. Variables included students' demographic and socio-economic background, high school characteristics, application timing, and personal intention. The students were divided into deciles based on their predicted enrollment probabilities, and the study suggested that it was more efficient to target the "fence-sitting" or middle groups rather in terms of yield likelihood than those with very high enrollment probabilities. Goenner and Pauls (2006) used logistic regression models with Bayesian model average techniques to predict the enrollment probabilities of 15,827 inquirers interested in attending the University of North Dakota in 2003 and to allocate recruitment efforts by geographic areas. They investigated the effects of inquiry contact methods, geographic factors, geodemographic factors, academic factors, and some interaction terms. The study suggested that recruitment efforts should be concentrated in geographic areas with high enrollment probabilities. Johnson (2019) examined the enrollment decisions of 42,950 out-of-state students admitted to a public research university from 2012 to 2016. He utilized mixed multinomial models and National Student Clearinghouse data to identify the students' destinations among five options: the study institution, another out-of-state public institution, an in-state public institution, a private institution, or a 2-year college. Factors included demographic characteristics, high school information, family background, financial aid offered, and admitted academic discipline. The study found that higher family incomes, being a family member of an alumnus, graduating from a feeder high school, being offered higher merit scholarships, or borrowing more loans increased students' likelihood of attending the institution. A surprising finding was that Pell-eligible students' enrollment decisions were not affected by grants in financial aid packages. Maldonado et al. (2017) used nested logit models to predict the enrollment decisions of 25,325 prospective students to four bachelor's programs in a small private Chilean university. The three possible outcomes were applied, admitted but not enrolled, and admitted and enrolled. The hierarchical models were deemed necessary because the latter two outcomes were more similar and should be grouped together. Factors included marketing efforts from the institution, students' socio-economic background, and stated preferences. The study found that on-campus activities and talks at secondary schools were more effective in encouraging enrollment than career fairs, male students were more likely to attend engineering and law programs, and students' online activities and stated preferences indicated their interests in attending the institution or individual programs. Braunstein et al. (1999) used logistic regression to model the enrollment decisions of 7,104 students admitted to Iona College in three academic years. The variables were grouped into three categories: demographic and social background, academic achievement and preparation, and financial aid. The study found that the demographic and social background did not affect students' enrollment decisions, but financial aid had a positive effect. The enrollment probability increased by 1.1% to 2.5% for every additional $1,000 offered, with loans having more influence than grants and work study having the least influence.

2.2.3 Conceptual Framework

Our conceptual framework for the deposit stage is rooted in the theory of college choice (Chapman 1979; Hossler, Braxton, and Coopersmith 1989) and prior studies on students' enrollment decisions. Students' decisions to accept admission offers are influenced by economic, sociological, and psychological factors (Paulsen 1990). According to the theory of human capital (Becker 2009), students evaluate the cost-benefit ratio of attending an institution based on economic factors. Sociological factors, such as sociological background (Johnson 2019) and status attainment (Hossler, Schmit, and Vesper 1999), also play a role in students' choice of institution. Psychological factors, including the institutional environment and climate, impact students' perceptions of student-institution fit (Paulsen 1990). Our hypothesis is that these factors not only influence whether students choose to pay deposits but also when they choose to pay. Admitted students are more likely to pay deposits sooner if they perceive the institution as a good investment, a source of status attainment, and/or a good fit for them.

With guidance from the theory of college choice and prior research, we categorize variables into three groups. The economic factors include offered financial aid and socio-economic status, i.e., Pell eligibility and expected family contribution

(EFC). Financial aid can impact students' enrollment decisions by reducing the cost of attendance, so we hypothesize that students are more likely to make a deposit when more financial aid is offered. We use the total financial aid offered instead of breaking it down by aid type, such as grant and scholarship, loan, and work study (Braunstein, McGrath, and Pescatrice 1999), because SFS suggests that the amounts of different aid types are correlated with each other. For example, the amount of federal loans depends on the grants and scholarships already offered.

The sociological factors include students' demographic characteristics, i.e., home distance from the university, gender, and racial ethnicity. These factors reflect the influence of parents, peers, counselors, and teachers on students' enrollment decisions (Johnson 2019). We hypothesize that these factors will have varying effects on students' deposit decisions.

Finally, psychological factors reflect students' desire to attend the institution (Paulsen 1990), and include admission to the Honors program, match between applied and admitted majors, attendance at recruitment events, and prompt review of admission decisions (Maldonado, Armelini, and Guevara 2017). Students who have a strong interest in the institution are more likely to pay deposits earlier, and we hypothesize that these interests are stronger when students are admitted to the Honors program, willing to be admitted to a different major, attend recruitment events, and promptly review their admission decisions.

2.3 Data and Variables

This study uses admission data of 60,285 admitted out-of-state students who intended to matriculate as first-time first-year students at the University of Delaware (Carnegie classification: R1), a public research university with an undergraduate population of around 18,000 students, during the Fall semesters of 2020, 2021, and 2022. The information was obtained from the Admissions Office and UD's enterprise data warehouse. To track the students' deposit decisions, the study period is defined from February 1 to the deposit deadline of May 1. February 1 is chosen as the starting point because the Admissions Office had made most admission decisions by then. The study period is further divided into eight intervals, February, March 1 to March 15, March 16 to March 31, April 1 to April 7, April 8 to April 14, April 15 to April 21, April 22 to April 28, and April 29 to May 1. The dataset is constructed as a student-period file with one row per student and period. Table 2.1 shows the number of observations and deposits by period each year. The number of observations increases from period 1 to either period 3 or period 4, reflecting the admission of students who were admitted after February. However, the number of observations decreases thereafter because students who paid deposits are no longer tracked. The number of deposits increases in April, especially after April 21, indicating the deadline effect on students' deposit decisions.

Table 2.1: Number of Observations and Deposits by Period Each Year

	2020		2021		2022	
Period	N	Deposit	N	Deposit	N	Deposit
1 - February	16,076	238	18,010	159	18,328	208
2 - March 1 to March 15	16,702	191	18,703	136	19,152	197
3 - March 16 to March 31	16,972	308	19,584	294	19,700	315
4 - April 1 to April 7	16,824	263	19,525	293	20,034	264
5 - April 8 to April 14	16,597	299	19,302	371	19,785	386
6 - April 15 to April 21	16,345	338	18,841	462	19,401	390
7 - April 22 to April 28	16,322	473	18,257	646	18,583	853
8 - April 29 to May 1	15,945	395	17,806	452	18,052	551

Table 2.2 describes the dependent and independent variables in the models. The dependent variable is whether a student paid a deposit by May 1. Of the 60,285 admitted students, 8,482 or 14.1% of made a deposit. The independent variables include the economic factors, sociological factors, and psychological factors. Three variables are continuous, Financial Aid, EFC, and Home Distance, with only Financial Aid having varying values by period. The rest of the variables are binary. On average, financial aid covers 28.6% of the Cost of Attendance (COA), and EFC covers 118.1% of COA. Of the 60,285 admitted students, 6,256 or 10.4% were eligible for Pell grants, 21,732 or 36.0% were male, 2,280 or 3.8% were African American, 3,457 or 5.7% were Asian, 5,072 or 8.4% were Hispanic, 44,164 or 73.3% were White, 1,813 or 3.0% were multi-ethnic, 17,478 or 29.0% attended early events for prospects, 6,998 or 11.6% were admitted to the Honors program, 2,913 or 4.8% were admitted to a different major than the one applied for, 3,443 or 5.7% visited campus, 3,127 or 5.2% attended the Decision Day event, and 24,449 or 40.6% did not review the admission decisions within two days.

Table 2.2: Description of the Dependent and Independent Variables

Variables	Description	N	Mean/Pct	S.D.
Dependent Variable				
Deposited	1 if deposited, 0 otherwise	8,482	14.1%	
Economic factors				
Financial Aid	Total offered financial aid over COA		0.286	0.162
EFC	Expected family contribution over COA		1.181	1.602
Pell	1 if Pell eligible, 0 otherwise	6,256	10.4%	
Sociological factors				

Home Distance	Home distance from UD		177.6	340.9
Male	1 if male, 0 if female	21,732	36.0%	
African American	1 if African American, 0 otherwise	2,280	3.8%	
Asian	1 if Asian American, 0 otherwise	3,457	5.7%	
Latino	1 if Latino/Hispanic/Chicano, 0 otherwise	5,072	8.4%	
White	1 if Caucasian, 0 otherwise	44,164	73.3%	
Multi-Ethic	1 if multi-ethnicity, 0 otherwise	1,813	3.0%	
Psychological factors				
Early Event	1 if attending early events for prospects	17,478	29.0%	
Honor	1 if admitted in the honors program, 0 otherwise	6,998	11.6%	
Major Change	1 if admitted major is different from applied major	2,913	4.8%	
Campus Tour	1 if attending campus tour, 0 otherwise	3,443	5.7%	
Decision Day	1 if attending Decision Day event, 0 otherwise	3,127	5.2%	
Delay Review	1 if not reviewing admission decision within 2 days	24,449	40.6%	

2.4 Statistical Model

We develop Bayesian hierarchical piecewise exponential models to analyze students' deposit decisions. These models, which are a type of discrete event history analysis, have a constant hazard function within each discrete time interval (Austin, 2017; DesJardins et al., 1994; Friedman, 1982). In this study, an event occurs if a student pays deposit between February 1 and May 1. Otherwise, the admitted student is "censored" on May 1 or "survives" from the desire to pay deposit. Equation (2.1) defines the logarithm of the hazard function $h_i[Period]$ to be the sum of baseline hazard $h_0[Period]$ and a linear combination of student attributes, where $x_{ij}[Period]$

represents the value of variable j for student i in a period, and $\beta_j[Period]$ represents the effect of the variable j in the period. The cumulative hazard, calculated as the product of the hazard function and the period length ($L[Period]$), is used to derive the logarithm of the survival function ($S_i[Period]$) in Equation (2.2), where $L[Period]$ is the number of days of each period. Then finally, Equation (2.3) calculates the probability of a student paying a deposit in a period ($\theta_i[Period]$) as one minus the survival function. In this model, the hazard function reflects the driving force behind a student's deposit decision, while the cumulative hazard represents the accumulated force over time. If a variable has a positive effect on the deposit decision, a higher value of that variable will increase the driving force and, as a result, increase the probability of the student paying a deposit, as described by the three equations.

$$log(h_i[Period]) = h_0[Period] + \sum_j \beta_j\,[Period]x_{ij}[Period] \tag{2.1}$$

$$log(S_i[Period]) = -h_i[Period]L[Period] \tag{2.2}$$

$$\theta_i[Period] = 1 - S_i[Period] \tag{2.3}$$

We introduce the Bayesian hierarchical framework to estimate the unknown coefficients $h_0[Period]$ and $\beta_j[Period]$ in Equation (2.1). Bayesian analysis models our initial uncertainty, or prior distribution, using probability distributions (McElreath, 2020). For the baseline hazard $h_0[Period]$, we model the initial uncertainty using normal distributions as shown in Equation (2.4). The $h^{MLE}[Period]$ is the maximum likelihood estimates (MLE) for $h_0[Period]$. We choose σ^0 to be 0.1 to make it a strong prior distribution. The prior distributions will be updated to posterior distributions using the observed data, and the data must show strong evidence for the posterior distributions to deviate from the prior distributions. The hierarchical structure is setup for estimating $\beta_j[Period]$. At the higher level, the variables have time-independent effects μ_j, i.e., the average effects over all periods. We assume μ_j are normally distributed with means being 0 as shown in Equation (2.5). This serves to reduce overfitting for the higher-level effects, because the observed data need to show enough support for non-zero parameter estimates. At the lower level, the variables have time-varying effects $\beta_j[Period]$. We assume $\beta_j[Period]$ are normally distributed with means being μ_j as shown in Equation (2.6). This also reduces overfitting, because the observed data need to show enough support for $\beta_j[Period]$ to deviate from μ_j. The standard deviations σ_h and σ control the uncertainty of these assumptions, and we model their uncertainty using exponential prior distributions with a rate parameter of 0.5 (λ = λ_h = 0.5 in Equations (2.7) and (2.8)). In summary, if the observed data support it, the posterior distributions of the higher-level effects will deviate from the zero-mean prior distributions, and the posterior distributions of the lower-level effects will deviate from the prior distributions centered at the higher-level effects, leading to variables with time-varying effects on deposit decisions.

$$h_0[Period] \sim \text{Normal}(h^{MLE}[Period], \sigma^0) \tag{2.4}$$

$$\mu_j \sim \text{Normal}(0, \sigma_h) \tag{2.5}$$

$$\beta_j[Period] \sim \text{Normal}\left(\mu_j, \sigma\right) \tag{2.6}$$

$$\sigma_h \sim \text{Exponential}(\lambda) \tag{2.7}$$

$$\sigma \sim \text{Exponential}(\lambda_h) \tag{2.8}$$

2.4.1 Constructing the Likelihood Function

Suppose a student i was observed in a period, the contribution to the likelihood function L_i depends on whether the student paid deposit in the period,

$$L_i = \begin{cases} S_i & \text{(No deposit)} \\ S_i h_i & \text{(Paid deposit)} \end{cases} \tag{2.9}$$

We can write the two cases in one equation. Let $y_i[Period]$ be the dependent variable, with 1 indicating a student paid deposit and 0 indicating the student did not pay. The likelihood function can be written as

$$L_i = h_i^{y_i}[Period]S_i[Period] \tag{2.10}$$

And the log-likelihood is

$$logL_i = y_i[Period]log(h_i[Period]) + log(S_i[Period])$$

According to Equations (2.2) and (2.1), it can be written as

$$\begin{aligned} logL_i &= y_i[Period]log(h_i[Period]) - h_i[Period]L[Period] \\ &= y_i[Period]\left(h_0[Period] + \sum_j \beta_j\,[Period]x_{ij}[Period]\right) \\ &\quad -exp\big(h_0[Period] + \textstyle\sum_j \beta_j\,[Period]x_{ij}[Period]\big)L[Period] \end{aligned} \tag{2.11}$$

Our dependent variable is imbalanced, because the out-of-state students' yield rate is below 15%. To address this challenge, we add a hyperparameter w to the log-likelihood to make the observations with deposit weight more than the others, so the log-likelihood becomes

$$\begin{aligned} logL_i = w * y_i[Period]\big(h_0[Period] + \textstyle\sum_j \beta_j\,[Period]x_{ij}[Period]\big) - \\ exp\big(h_0[Period] + \textstyle\sum_j \beta_j\,[Period]x_{ij}[Period]\big)L[Period] \end{aligned} \tag{2.12}$$

This log-likelihood function defined above is used to estimate the $h^{MLE}[Period]$ in equation (2.4) and to construct the posterior distribution below.

2.4.2 Constructing the Posterior Distribution

The posterior distribution is proportional to the product of likelihood function and prior distribution, i.e.,

$$p\big(y_i[Period] | h_0[Period], \beta_j[Period], x_{ij}[Period]\big)$$
$$\propto L_i * p(h_0[Period]) * p\big(\beta_j[Period]\big) \tag{2.13}$$

From Equation (2.4), we have

$$p(h_0[Period]) = p(h_0[Period] | h^{MLE}[Period], \sigma^0)$$
$$= pdf\big(\text{Normal}(h_0[Period] | h^{MLE}[Period], \sigma^0)\big) \tag{2.14}$$

Where pdf represents the probability density function of a distribution.

In a Bayesian hierarchical model, the prior distributions of the parameters depend on their own prior distributions, i.e., the distributions of the higher-level parameters. From Equation (2.5) to Equation (2.8), we have

$$p\big(\beta_j[Period]\big) = p\big(\beta_j[Period] | \mu_j, \sigma\big)$$
$$= pdf\left(\text{Normal}\big(\beta_j[Period] | \mu_j, \sigma\big)\right) * p\big(\mu_j\big) * p(\sigma) \tag{2.15}$$

$$p\big(\mu_j\big) = pdf\left(\text{Normal}\big(\mu_j | 0, \sigma_h\big)\right) * p(\sigma_h) \tag{2.16}$$

$$p(\sigma) = pdf\big(\text{Exponential}(\sigma | \lambda)\big) \tag{2.17}$$

$$p(\sigma_h) = pdf\big(\text{Exponential}(\sigma_h | \lambda_h)\big) \tag{2.18}$$

Input Equations from (2.15) to (2.18) into Equation (2.13) and take the logarithm of $p\big(y_i[Period] | h_0[Period], \beta_j[Period], x_{ij}[Period]\big)$, we have

$$log\left(p\big(y_i[Period] | h_0[Period], \beta_j[Period], x_{ij}[Period]\big)\right)$$
$$\propto w * y_i[Period]\left(h_0[Period] + \sum_j \beta_j\,[Period] x_{ij}[Period]\right)$$

$$
\begin{aligned}
&-exp\big(h_0[Period] + \textstyle\sum_j \beta_j\,[Period]x_{ij}[Period]\big)L_t \\
&+logpdf\big(\text{Normal}(h_0[Period]|h^{MLE}[Period],\sigma^0)\big) \\
&+logpdf\,\Big(\text{Normal}\big(\beta_j[Period]|\mu_j,\sigma\big)\Big) \\
&+logpdf\,\Big(\text{Normal}\big(\mu_j|0,\sigma_h\big)\Big) \\
&+logpdf\big(\text{Exponential}(\sigma|\lambda)\big) \\
&+logpdf\big(\text{Exponential}(\sigma|\lambda_h)\big) \qquad (2.19)
\end{aligned}
$$

2.4.3 Parameter Estimation

The MLE results from the piecewise exponential models without the Bayesian hierarchical structure are obtained using the Ipopt solver in the JuMP package (v1.10.0) (Lubin et al., 2023) in Julia (v1.7.2). The MLE results are not only used for $h^{MLE}[Period]$ in Equation (2.4), but also used as the starting values in the training of the Bayesian hierarchical piecewise expoential models. We run three Markov chains to estimate the posterior distributions for the Bayesian hierarchical piecewise expoential models using the DynamicHMC package, each consisting of 1,000 samples after a series of warm-up steps to find an appropriate step size for the "No-U-Turn Sampler" (NUTS) (Betancourt, 2017; Hoffman et al., 2014). The gradients of the log densities for NUTS are obtained from automatic differentiation using the Zygote package (v0.6.60), and the kinetic energy is the default Gaussian with identity matrix. Appendix A contains Julia code for the construction and training of the Bayesian hierarchical piecewise exponential models.

2.5 Results and Discussion

We derive the point estimates and credible intervals from the posterior distributions. The results are interpreted based on the credible intervals of estimated coefficients and the corresponding hazard functions. If all credible intervals span across zero, the variable is considered not important in the deposit decision. A variable is defined to have a time-varying effect if the credible intervals do not overlap in at least two periods. Otherwise, the variable is defined to have a time-independent effect. The effect of a variable is measured by the hazard ratio, which is the exponential of the product of the variable change and its estimated coefficient, holding other variables and their coefficients constant according to Equation (2.1). The hazard ratio for the baseline hazard is the exponential of the difference between the estimated hazards in two periods. For binary variables, the ratio is the exponential of the estimated coefficient in the period, with a positive relationship between the ratio and the estimated coefficient. A variable has a positive effect on the deposit decision if the hazard ratio is larger than 1 or the estimated coefficient is larger than 0, and vice versa.

2.5.1 Fall 2020

Table 2.3 shows the lower bounds of 95% credible intervals, the means, and the upper bounds of 95% credible intervals of parameter estimates for students who were admitted for Fall 2020, including the baseline hazard and the variables. The baseline hazard differs among periods. For example, the 95% credible interval (-8.22, -7.88) in first period does not overlap with interval (-5.98, -5.65) in the last period. The baseline hazard shows a fluctuating increasing trend from period 1 to period 8. The largest increase is between period 6 and period 7, which implies that students experience deadline pressure in the fourth week of April. The hazard ratio is 9.30 between the first and the last periods, indicating a much higher probability of deposit payment in the last period than in the first period.

Table 2.3: Parameter Estimates for 2020, 2.5% percentile, mean and 97.5% percentile

Variable	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7	Period 8
Baseline	-8.22, -8.05, -7.88	-8.55, -8.37, -8.19	-7.68, -7.51, -7.34	-7.45, -7.27, -7.1	-7.09, -6.91, -6.74	-7.31, -7.14, -6.97	-6.14, -5.98, -5.83	-5.98, -5.82, -5.65
Financial Aid	-0.9, -0.47, -0.09	-0.7, -0.23, 0.21	-0.54, -0.13, 0.28	-0.3, 0.07, 0.46	-0.48, -0.07, 0.37	-0.3, 0.12, 0.54	-0.59, -0.21, 0.16	-0.38, 0.02, 0.41
Pell	-0.27, 0.09, 0.41	-0.17, 0.19, 0.52	-0.02, 0.24, 0.51	-0.17, 0.15, 0.44	-0.11, 0.19, 0.47	-0.12, 0.16, 0.45	-0.04, 0.19, 0.43	0.29, 0.55, 0.8
EFC	-0.23, -0.13, -0.04	-0.12, -0.03, 0.06	-0.06, 0.0, 0.06	-0.13, -0.05, 0.03	-0.18, -0.09, -0.01	-0.13, -0.05, 0.01	-0.18, -0.11, -0.04	-0.14, -0.06, 0.01
Home Distance	-0.26, -0.06, 0.12	-0.6, -0.29, -0.02	-0.22, -0.06, 0.08	-0.12, 0.02, 0.15	-0.37, -0.17, -0.0	-0.19, -0.04, 0.09	-0.21, -0.08, 0.03	-0.23, -0.09, 0.04
Gender	-0.56, -0.34, -0.11	-0.3, -0.07, 0.15	-0.52, -0.31, -0.11	-0.42, -0.2, 0.01	-0.26, -0.06, 0.13	-0.0, 0.2, 0.38	-0.12, 0.03, 0.18	0.23, 0.41, 0.59
Asian	-0.65, -0.21, 0.17	-0.6, -0.14, 0.31	-0.6, -0.17, 0.2	-0.42, -0.0, 0.4	-0.58, -0.17, 0.21	-0.52, -0.1, 0.31	-0.59, -0.22, 0.12	-0.44, -0.08, 0.28
African American	-0.47, -0.02, 0.4	-0.48, -0.01, 0.44	-0.46, -0.06, 0.36	-0.26, 0.17, 0.61	-0.32, 0.09, 0.49	-0.24, 0.2, 0.6	-0.37, 0.01, 0.35	-0.14, 0.26, 0.61
Hispanic	-0.22, 0.13, 0.47	0.02, 0.37, 0.75	-0.08, 0.22, 0.53	-0.0, 0.35, 0.7	-0.21, 0.14, 0.49	-0.0, 0.31, 0.62	-0.17, 0.12, 0.42	0.02, 0.32, 0.62
White	-0.2, 0.07, 0.32	0.23, 0.48, 0.76	-0.09, 0.14, 0.38	0.1, 0.35, 0.61	0.25, 0.48, 0.71	0.15, 0.38, 0.59	-0.19, 0.01, 0.21	0.1, 0.31, 0.53
Multi-	-0.48, -	-0.44,	-0.32,	-0.16,	-0.41, -	-0.29,	-0.14,	-0.02,

Ethic	0.02, 0.4	0.02, 0.46	0.09, 0.45	0.27, 0.68	0.01, 0.41	0.12, 0.52	0.2, 0.57	0.39, 0.78
Early Event	0.74, 0.96, 1.17	0.72, 0.95, 1.18	0.52, 0.71, 0.91	0.61, 0.82, 1.03	0.39, 0.59, 0.79	0.7, 0.88, 1.07	0.36, 0.52, 0.68	0.12, 0.31, 0.51
Honor Program	-0.34, -0.06, 0.18	-0.36, -0.05, 0.24	-0.1, 0.15, 0.38	0.07, 0.33, 0.57	-0.21, 0.07, 0.34	-0.21, 0.03, 0.27	0.0, 0.21, 0.4	-0.1, 0.14, 0.37
Change Major	0.75, 1.05, 1.33	0.66, 0.98, 1.33	0.82, 1.08, 1.37	0.75, 1.05, 1.35	0.79, 1.08, 1.36	0.92, 1.16, 1.41	0.82, 1.08, 1.32	0.83, 1.12, 1.41
Campus Tour	1.59, 1.9, 2.2	1.36, 1.69, 2.01	1.5, 1.72, 1.94	1.61, 1.84, 2.05	1.46, 1.69, 1.9	1.55, 1.75, 1.96	1.52, 1.71, 1.89	1.48, 1.68, 1.88
Decision Day Event	1.56, 1.78, 2.0	1.59, 1.84, 2.09	1.39, 1.6, 1.82	1.17, 1.41, 1.68	1.21, 1.45, 1.7	1.44, 1.67, 1.9	1.11, 1.33, 1.55	1.07, 1.34, 1.58
Delay Review	-0.62, -0.38, -0.16	-0.49, -0.25, 0.01	-0.51, -0.29, -0.09	-0.29, -0.08, 0.13	-0.37, -0.14, 0.07	-0.3, -0.11, 0.09	-0.17, -0.0, 0.15	-0.09, 0.08, 0.24

Five of the sixteen variables have time-varying effects, and they are Gender, White, Early Event, Decision Day, and Delay Review. The 95% credible intervals of Gender do not overlap between the first (-0.58, -0.04) period and the last period (-0.01, 0.32), indicating the effects are different in the two periods. The parameter estimates of Gender show an increasing trend, indicating the female students tend to pay deposit earlier than the male students. For example, the hazard ratio is 0.73 in the first period, indicating female students are more likely to pay deposits in the period. However, the hazard ratio changes to 1.16, indicating male students are more likely to pay in the last period. This finding would be neglected, if we assume variables to have time-independent effects. The point estimate for Gender would be close to 0 at -0.0079, indicating gender plays little role for deposit decisions. The parameter estimates of Delay Review also show an increasing trend. Students tend not to pay deposits in early periods, if they postpone to review admission decisions for at least two days, but the delay does not matter in later periods. In contrast, the parameter estimates are positive but with a declining trend for early events and decision day event, indicating that the encouragement from attending the events fades over time. The parameter estimates of White fluctuate among periods. White students are more likely to pay deposits in some periods but not in the other periods.

Eight variables have time-independent effects on deposit decisions, and they are Financial Aid, Pell, EFC, Home Distance, Hispanic, Honors Program, Major Change, and Campus Tour, because all corresponding 95% credible intervals overlap with each other. Financial aid does not matter except in the first period. It is not surprising in 2020, because students' decisions are more affected by Covid-19 than financial burdens. In contrast, Pell eligibility does not matter until the last period. The hazard ratio is 1.45 in the last period, indicating Pell eligible students are more likely to pay deposits than non-Pell eligible students. This makes sense to the Admissions Office, because Pell eligible students would like to delay any financial expense until they cannot. This also suggests to them that the recruitment effort for Pell students may not appear effective until the last period. Other variables have important effects in more than one period. The parameter estimates of EFC and Home Distance are negative in some periods, indicating students with lower income and those closer to UD are more likely to pay deposits. On the other hand, the parameter estimates of Hispanic and Honor Program are positive in some periods, indicating Hispanic students and students who are admitted to the Honors program are more likely to pay deposits. Lastly, the parameter estimates of Major Change and Campus Tour are positive in all periods, indicating students are consistently more likely to pay deposits if they are willing to be admitted a major different than an applied major, or they attend campus tours.

Three variables do not have important effect on deposit decisions in any period, and they are Asian, African American, and Multi-Ethnic. The credible intervals of the parameter estimates span across zero in all periods, indicating the three factors do not matter when making deposit decisions.

2.5.2 Fall 2021

Table 2.4 shows the lower bounds of 95% credible intervals, the means, and the upper bounds of 95% credible intervals of parameter estimates for students who were admitted for Fall 2021, including the baseline hazard and the variables. Similar to 2020, the baseline hazard varies across periods, with non-overlapping 95% credible intervals. For instance, the first period has an interval of (-8.39, -8.04), while the last period has an interval of (-6.10, -5.78). The baseline hazard exhibits an increasing trend, indicating an increasing likelihood of deposit payments over time. The hazard ratio is 9.68 between the first and the last periods, indicating students are much more likely to pay deposits in the last period than the first period.

Table 2.4: Parameter Estimates for 2021, 2.5% percentile, mean and 97.5% percentile

Variable	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7	Period 8
Baseline	-8.39, -8.21, -8.04	-7.72, -7.55, -7.38	-7.69, -7.51, -7.34	-6.9, -6.74, -6.57	-6.6, -6.43, -6.26	-6.74, -6.59, -6.42	-6.64, -6.49, -6.33	-6.1, -5.94, -5.78
Financial Aid	0.54, 1.03, 1.52	0.64, 1.13, 1.62	0.71, 1.12, 1.53	0.85, 1.26, 1.68	0.82, 1.24, 1.63	1.08, 1.49, 1.88	1.34, 1.7, 2.08	1.12, 1.5, 1.9
Pell	-0.64,	-0.78,	-0.42,	-0.71,	-0.49,	-0.58,	-0.33,	-0.06,

	-0.25, 0.12	-0.4, -0.03	-0.11, 0.21	-0.37, -0.06	-0.17, 0.12	-0.32, -0.04	-0.09, 0.15	0.19, 0.44
EFC	-0.15, -0.05, 0.04	-0.33, -0.18, -0.04	-0.04, 0.03, 0.09	-0.11, -0.03, 0.05	-0.12, -0.04, 0.03	-0.15, -0.07, -0.0	-0.05, -0.0, 0.04	-0.13, -0.05, 0.01
Home Distance	-0.3, -0.09, 0.07	-0.41, -0.17, 0.03	-0.31, -0.13, 0.0	-0.07, 0.04, 0.13	-0.27, -0.12, 0.01	-0.38, -0.22, -0.08	-0.26, -0.15, -0.04	-0.21, -0.09, 0.02
Gender	-0.47, -0.21, 0.03	-0.53, -0.25, 0.02	-0.45, -0.25, -0.05	-0.39, -0.18, 0.04	-0.4, -0.2, 0.0	-0.25, -0.08, 0.11	0.09, 0.23, 0.37	0.15, 0.32, 0.49
Asian	-0.86, -0.39, 0.04	-0.96, -0.49, -0.06	-0.99, -0.56, -0.13	-0.85, -0.41, -0.02	-0.87, -0.46, -0.06	-0.68, -0.29, 0.07	-0.61, -0.27, 0.06	-0.77, -0.38, 0.01
African American	-0.85, -0.35, 0.11	-0.68, -0.24, 0.19	-0.64, -0.23, 0.15	-0.61, -0.2, 0.25	-0.92, -0.48, -0.06	-0.76, -0.37, -0.02	-0.39, -0.05, 0.28	-0.5, -0.12, 0.25
Hispanic	-0.42, -0.03, 0.34	-0.32, 0.04, 0.39	-0.32, -0.0, 0.31	-0.23, 0.09, 0.41	-0.22, 0.08, 0.36	0.02, 0.29, 0.58	0.03, 0.28, 0.54	-0.1, 0.18, 0.46
White	-0.3, -0.03, 0.22	-0.41, -0.16, 0.1	-0.21, 0.02, 0.25	-0.17, 0.03, 0.26	-0.23, -0.01, 0.2	-0.19, 0.02, 0.21	-0.1, 0.07, 0.26	-0.03, 0.18, 0.41
Multi-Ethic	-0.54, -0.11, 0.33	-0.62, -0.15, 0.26	-0.37, 0.04, 0.42	-0.53, -0.13, 0.29	-0.3, 0.11, 0.47	-0.28, 0.13, 0.49	-0.61, -0.22, 0.15	-0.36, 0.02, 0.4
Early Event	0.09, 0.34, 0.57	0.07, 0.33, 0.61	0.11, 0.3, 0.51	0.3, 0.51, 0.74	0.03, 0.23, 0.42	-0.06, 0.13, 0.33	-0.14, 0.03, 0.18	-0.38, -0.2, -0.01
Honor Program	-0.85, -0.5, -0.15	-0.9, -0.51, -0.18	-0.71, -0.42, -0.13	-0.61, -0.32, -0.06	-0.56, -0.3, -0.05	-0.36, -0.12, 0.1	-0.33, -0.12, 0.08	-0.33, -0.07, 0.16
Change Major	0.63, 0.96, 1.28	0.24, 0.58, 0.91	0.39, 0.66, 0.93	0.45, 0.71, 0.98	0.51, 0.76, 0.99	0.55, 0.76, 0.98	0.38, 0.61, 0.82	0.57, 0.82, 1.04
Campus Tour	1.25, 1.6, 1.96	1.31, 1.62, 1.95	1.2, 1.46, 1.7	1.3, 1.52, 1.76	1.19, 1.4, 1.6	1.2, 1.41, 1.59	1.24, 1.42, 1.59	1.51, 1.7, 1.91
Decision Day Event	1.03, 1.36, 1.66	0.74, 1.09, 1.43	0.91, 1.2, 1.5	0.59, 0.93, 1.23	0.72, 0.99, 1.26	1.25, 1.45, 1.66	1.37, 1.55, 1.72	1.09, 1.31, 1.52
Delay	-0.66,	-0.77,	-0.3,	-0.41,	-0.29,	-0.11,	-0.08,	-0.06,

Review	-0.39, -0.13	-0.46, -0.2	-0.12, 0.07	-0.21, 0.0	-0.1, 0.09	0.06, 0.24	0.06, 0.2	0.1, 0.26

Five variables exhibit time-varying effects on deposit decisions: Home Distance, Gender, Early Event, Decision Day, and Delay Review. Four of these variables, Home Distance, Gender, Early Event, and Delay Review, show similar patterns with the previous year. Specifically, students who live closer to UD are more likely to pay deposits, female students tend to pay earlier than males, the impact of attending early events wanes over time, and students who delay reviewing their admissions decision tend not to pay deposits in the first two periods. However, the variable Decision Day exhibits a new pattern: while attending Decision Day events remains a positive factor for deposit decisions, its effect does not show a decreasing trend. The hazard ratio decreases from 3.90 in period 1 to 2.53 in period 4, before increasing to 4.71 in period 7. This is due to UD hosting another Decision Day event in April, which causes the effect to fade after February and then increase in April.

Most of other variables have time-independent effects, except for White and Multi-Ethnic, and they are Financial Aid, Pell, EFC, Asian, African American, Hispanic, Honors Program, Major Change, and Campus Tour. Unlike the previous year, students are incentivized by financial aid in all periods, which suggests that financial burden had been a more pressing concern than COVID-19. For instance, in period 7, a 10% increase in financial aid, or about $5.3K with a cost of attendance of $53,422 in 2021, results in a hazard ratio of 1.19. Assuming a student has a 15% chance of paying the deposit, their likelihood increases to 17.5% with the additional 10% financial aid, assuming other factors remain constant. Honors Program also exhibits a different pattern, with parameter estimates turning negative in most periods, indicating that students are less likely to pay deposits if admitted to the Honors

Program. This change could be due to the change of admission policies, according to the Admissions Office. Furthermore, Asian and African American have become important factors in 2021, and students belonging to these racial-ethnic groups are less likely to pay deposits in some periods. EFC, Hispanic, Change Major, and Campus Tour continue to have similar effects as in the previous year, i.e., students are more likely to pay deposits if they come from lower-income families, are Hispanic students, have been admitted to a major different from their applied major, or have attended campus tours.

2.5.3 Fall 2022

Table 2.5 shows the lower bounds of 95% credible intervals, the means, and the upper bounds of 95% credible intervals of parameter estimates for students who were admitted for Fall 2022, including the baseline hazard and the variables. The baseline hazard still shows an increasing trend, and students are mostly likely to pay deposit in the last period. The hazard ratio between the first and the last periods is 74.44, which is much higher than the previous two years. This is mostly due to the much lower baseline hazard in the first period, i.e., $h_0[1]$ is -9.79 in 2022 vs -8.21 and -8.05 in 2021 and 2020, respectively.

Table 2.5: Parameter Estimates for 2022, 2.5% percentile, mean and 97.5% percentile

Variable	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7	Period 8
Baseline	-9.98, -9.79, -9.61	-8.42, -8.24, -8.07	-7.77, -7.58, -7.41	-6.76, -6.59, -6.4	-7.49, -7.3, -7.13	-7.55, -7.39, -7.21	-6.49, -6.33, -6.16	-5.65, -5.48, -5.32
Financial Aid	0.23, 0.73, 1.25	-0.22, 0.29, 0.77	-0.04, 0.41, 0.89	-0.27, 0.2, 0.63	0.02, 0.5, 0.96	0.26, 0.7, 1.13	0.24, 0.62, 0.99	-0.03, 0.41, 0.85

Pell	-0.06, 0.32, 0.65	-0.26, 0.09, 0.44	-0.38, -0.06, 0.25	-0.34, 0.01, 0.32	-0.02, 0.29, 0.57	-0.16, 0.15, 0.44	-0.08, 0.14, 0.36	0.15, 0.37, 0.58
EFC	-0.04, 0.03, 0.09	-0.09, -0.0, 0.07	-0.1, -0.03, 0.05	-0.07, 0.0, 0.06	-0.09, -0.02, 0.04	-0.08, -0.02, 0.05	-0.1, -0.04, 0.01	-0.25, -0.17, -0.09
Home Distance	-0.55, -0.26, -0.02	-0.63, -0.35, -0.1	-0.3, -0.12, 0.03	-0.19, -0.04, 0.09	-0.42, -0.23, -0.04	-0.29, -0.13, 0.01	-0.11, -0.02, 0.06	-0.22, -0.09, 0.03
Gender	-0.58, -0.31, -0.04	-0.48, -0.22, 0.01	-0.32, -0.1, 0.1	-0.33, -0.11, 0.11	-0.27, -0.08, 0.1	-0.14, 0.06, 0.24	-0.15, -0.03, 0.09	-0.01, 0.15, 0.32
Asian	-0.63, -0.09, 0.4	-0.47, -0.0, 0.49	-0.53, -0.11, 0.31	-0.56, -0.16, 0.23	-0.3, 0.14, 0.58	-0.07, 0.31, 0.68	-0.57, -0.24, 0.1	-0.52, -0.13, 0.24
African America n	-0.38, 0.17, 0.7	-0.57, -0.06, 0.44	-0.68, -0.22, 0.23	-0.68, -0.21, 0.27	-0.58, -0.1, 0.34	-0.31, 0.12, 0.56	-0.23, 0.09, 0.42	-0.49, -0.11, 0.24
Hispanic	0.06, 0.48, 0.89	-0.14, 0.26, 0.65	-0.24, 0.1, 0.45	-0.21, 0.13, 0.48	-0.05, 0.28, 0.62	0.07, 0.41, 0.74	-0.13, 0.12, 0.35	-0.1, 0.19, 0.47
White	0.67, 0.95, 1.25	0.09, 0.34, 0.62	-0.12, 0.13, 0.37	-0.38, -0.15, 0.1	0.24, 0.46, 0.72	0.14, 0.36, 0.6	-0.17, 0.03, 0.24	-0.23, -0.01, 0.2
Multi-Ethic	-0.57, -0.01, 0.55	-0.86, -0.26, 0.28	-0.67, -0.15, 0.32	-0.92, -0.37, 0.18	-0.64, -0.1, 0.44	-0.34, 0.17, 0.64	-0.51, -0.08, 0.31	-0.51, -0.07, 0.39
Early Event	0.86, 1.11, 1.37	0.55, 0.78, 1.03	0.59, 0.79, 0.98	0.62, 0.84, 1.07	0.38, 0.58, 0.75	0.53, 0.72, 0.91	0.28, 0.43, 0.59	0.03, 0.21, 0.38
Honor Program	-0.77, -0.45, -0.14	-0.74, -0.4, -0.09	-0.52, -0.25, 0.03	-0.31, -0.02, 0.27	-0.36, -0.11, 0.13	-0.31, -0.04, 0.21	-0.27, -0.1, 0.09	-0.36, -0.13, 0.1
Change Major	0.97, 1.28, 1.6	0.56, 0.91, 1.25	0.8, 1.07, 1.36	0.57, 0.9, 1.2	0.57, 0.85, 1.12	0.64, 0.92, 1.18	0.55, 0.74, 0.92	0.86, 1.1, 1.35
Campus Tour	1.42, 1.71, 2.0	1.49, 1.75, 2.03	1.27, 1.53, 1.78	1.14, 1.39, 1.67	1.35, 1.55, 1.76	1.65, 1.84, 2.05	1.51, 1.65, 1.79	1.54, 1.72, 1.89
Decision Day Event	1.63, 1.89, 2.14	1.47, 1.75, 2.03	1.34, 1.6, 1.84	1.19, 1.5, 1.78	2.19, 2.39, 2.59	1.73, 1.94, 2.14	2.24, 2.37, 2.5	1.99, 2.15, 2.32

Delay Review	-0.45, -0.17, 0.07	-0.35, -0.1, 0.13	-0.36, -0.17, 0.02	-0.54, -0.31, -0.09	-0.42, -0.22, -0.03	-0.13, 0.07, 0.26	0.1, 0.22, 0.35	0.12, 0.27, 0.42

Seven variables exhibit time-varying effects: EFC, Gender, White, Early Events, Change Major, Decision Day, and Delay Review. The patterns for EFC, Gender, Early Events, and Decision Day remain consistent with the previous year. That is to say, female students tend to pay deposits earlier than male students, the encouragement from attend early events fades over time, and students are encouraged to pay deposits from individual Decision Day events. EFC and Change Major become variables with time-varying effects, but their general patterns do not change much from the previous years. Lower-income students are more likely to pay deposits in the last period, and students are more likely to pay deposits when admitted majors differ from their applied majors. Similar to 2020, White students are more likely to pay deposits in some periods and exhibit similar behavior to non-White students in other periods.

Six variables have time-independent effects: Financial Aid, Pell, Home Distance, Hispanic, Honors Program, and Campus Tour. Financial Aid remains as a positive factor, although only in half of the periods, and with smaller hazard ratios than in 2021. For example, the hazard ratio decreases from 1.19 in 2021 to 1.06 in 2022 for period 7. With the same assumption with 2021, an additional 10% in financial aid would increase the chance of paying the deposit from 15% to 15.9%. Pell-eligible students exhibit a similar behavior to that of 2020, with a higher likelihood of paying deposits in the last period. The other variables have similar patterns to the previous year. That is, students are more likely to pay if they live closer to UD, are Hispanic, and attend campus tours, but being admitted to the Honors program does not encourage early deposit payment.

Three variables do not have important impact on deposit decisions, and they are Asian, African American, and Multi-Ethnic. The list is the same with 2020.

2.5.4 Comparison Among the Three Years

To compare the results among the three years, we focused on the baseline hazard and one variable from each of the three factor groups: Financial Aid from economic factors, Gender from sociological factors, and Decision Day from psychological factors. The parameter estimate distributions for each variable are shown in Figure 2.1 as boxplots. The baseline hazard for all three years shows an increasing trend, with students being less likely to pay deposits in the early periods and more likely in the latter periods. This is within expectation, because students would like to compare institutions' admission offers but do not want to miss the deposit deadline. However, the growth path of the baseline hazard differs among the years. In 2020, the hazard slowly increases in the first six periods, jumps in period 7, and the difference between period 7 and period 8 is not large. In 2021, the hazard slowly increases in the first three periods, jumps in period 4, and remains relatively stable until the jump in the last period. In 2022, the hazard is much lower in the first period, catches up and fluctuates in the following several periods, and jumps to the strongest in the last period. The uncertainty of the baseline hazard makes it difficult for the Admissions Office to decide how many students to admit after February. For example, the first six periods suggest that the yield will be low in 2022, so more students need to be admitted to hit admission targets, but the last two periods, especially the last period, do not support the decision.

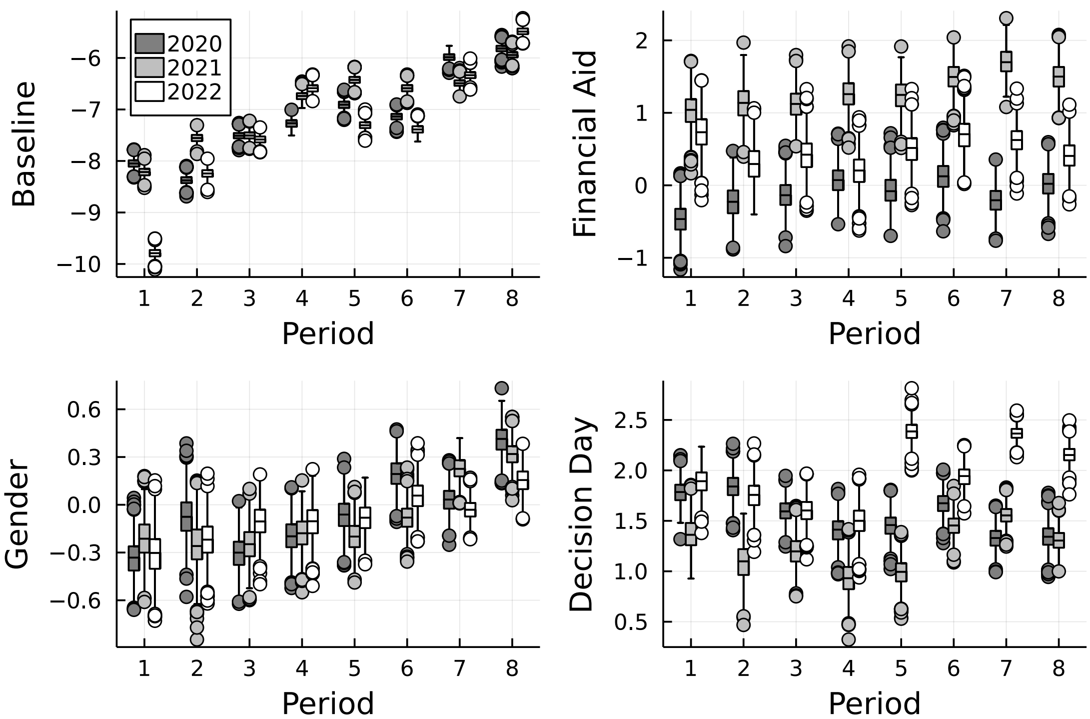

Figure 2.1: Boxplots of the distributions of the parameter estimates of baseline force, Financial Aid, Gender, and Decision Day among the three years.

Financial aid has had a varying impact on deposit decisions over the past three years. In 2020, the boxplots locate close to zeros, indicating financial aid does not affect students' deposit decisions after February. Financial aid is supposed to reduce students' cost of attendance, but financial burden was not what students paid attention to due to Covid-19. However, as the pandemic-related concerns decreased in 2021, the boxplots move up above zero, indicating that financial aid became a positive factor in encouraging students to accept admission offers. Institutional grants and scholarships are the primary source of financial aid offered, and it is encouraging for the admissions office to see that students value and respond positively to the institution's financial support. For 2022, the boxplots barely overlap with those in 2021 except the first period, and they move closer to zero, indicating students have weaker response to financial aid in 2022 than 2021. This presents a challenge to the SFS team responsible for packaging students' financial aid. If the budget for institutional aid remains the same, but its impact on deposit decisions is not stable year over year, how can the team work more effectively with the admissions office to optimize the distribution of grants and scholarships to increase yield? We observe that the effect of financial aid has been relatively stable within a year, so one possible solution is to monitor the effect of financial aid on deposit decisions early, and adjust the financial aid policies promptly, or adjust the expectation of the financial aid to increase deposits.

Unlike financial aid, gender has consistent effects across the three years in terms of the general pattern, but the effects vary by period within each year. Female students tend to pay deposits earlier than male students, as shown by the rising boxplots over periods in all three years. In the first few periods, the boxplots are mostly below zero, indicating that male students are less likely to pay deposits early on. The boxplots then gradually rise and are mostly above zero in the last periods, suggesting that male students are more likely to pay deposits towards the end of the deposit period. That is to say, the deadline effect plays a more important role to male students than females, or female students are more eager to accept UD's admission offers. We suspect this is due to the larger population of female students at UD, e.g., 59% of undergraduate students pursing bachelor's degrees are female at UD in Fall 2021. For comparison, it is 47% and 49% in Penn State University and University of Maryland in Fall 2021, respectively. According to the College Choice theory, students prefer an institution with matching sociological pattern, so female students are more likely consider UD as a good fit and thus are more willing to pay deposits early. In a conversation with an admissions professional that works with the data, he was very interested in this finding, but it would be overlooked if we assume that factors do not have time-varying effects. Models that make this assumption estimate the average effect across periods and would suggest that gender has no impact on deposit decisions, as the average parameter estimates would be close to zero, as seen in the boxplots.

The Admissions Office's Decision Day event has been effective in encouraging admitted students to pay their deposits, as indicated by the boxplots consistently being above zero in all periods across the three years. Moreover, the boxplots reflect stimulation from individual events from Februarys and Aprils. The boxplots also reveal the impact of individual events held in February and April. In 2020, due to Covid-19, only one Decision Day event was held in February and those in April were canceled. The boxplots suggest that the effect of the event diminishes over time, indicating that while the event may initially increase students' desire to pay their deposits, the effect decreases if they do not do so shortly after the event. In 2021, UD has multiple Decision Day events in February and April. The boxplots indicate that the effect fades after February but increase again in April. In 2021, multiple Decision Day events were held in both February and April, and the boxplots show a decrease in effect after February followed by an increase after the April event. This trend is more pronounced in 2022, where there is a large jump in effect after the April event following a decrease between February and April.

Other variables are briefly discussed here. In 2020 and 2022, Pell eligibility does not affect the deposit decisions until the last period when Pell eligible students are more likely to pay deposits, but this does not apply to 2021. EFC does not matter in most periods in the three years, but the negative parameter estimates in a few periods indicate that students in higher income families are less likely to pay deposits. Home distance had a negative effect on deposit decisions in all three years, indicating that students living closer to UD were more likely to pay deposits. In terms of racial ethnicity, Hispanic students are more likely to pay deposits in at least two periods each year, Multi-Ethnic students do not show preference in all three years, Asian and African American students are less likely to pay deposits in some periods in 2021 but they do not show preference in 2020 and 2022, and White students are more likely to pay deposits in half of the periods in 2020 and 2022 but they do not show preference in 2021. Students who attended early events are more likely to pay deposits in all three years, but the effect tend to be stronger in 2020 and 2022 than 2021. Students who are admitted to the Honors program are more likely to pay deposits in some periods in 2020, but it becomes a negative factor in 2021 and 2022. Students with admitted majors different than applied majors are more likely to pay deposits in all three years. Campus tours encourage students to pay deposits in all three years. Lastly, students who delay to review admission decisions are less likely to pay deposits in the early periods, but this effect fades over time, particularly in 2022.

2.6 Conclusion

This chapter provides valuable insights for an institution to better understand admitted students' deposit decisions and adjust their recruitment strategies accordingly, facilitating a better budgetary and student life process. While this study is limited to one institution for out-of-state students who intended to matriculate between Fall 2020 and Fall 2022, the findings shed light on the time-varying effects of various factors on students' deposit decisions, which can be useful for other institutions in similar contexts. The results from the Bayesian hierarchical piecewise exponential models validate our hypothesis that students behave or respond to recruitment efforts differently in different periods during the admission season. For example, female students are more likely to pay deposits in early periods than male students, while in later periods male students become more likely to pay. This finding would be neglected if we assume time-independent effects in the event history analysis. Moreover, Pell eligible students are most likely to pay deposits in the last period, because they would like to delay any financial expense as much as possible. The Admissions Office could potentially help these students pay earlier by designing policies such as partially waiving their deposits. Regarding campus events, the findings show that students do respond to early events for prospects and Decision Day events for admitted students. These events act as impulse forces that encourage students to pay deposits, and their effects decrease over time. Therefore, the encouragement from early events fades over time, since they all happen before the first period, and the strongest effects of Decision Day events are observed in the periods when the events occur. These findings confirm the Admissions Office's efforts to help students recognize UD as an excellent institution for their undergraduate study.

It's important to note that students' behaviors can change over the course of the three-year study period, influenced by factors such as macro environment and admission policies. For example, financial aid changes from an irrelevant factor to positive factor from 2020 to 2021 and 2022, likely due to the change of students' concern over Covid-19. Another surprising finding was that being admitted to the Honors program changed from an encouragement to a discouragement factor from 2020 to 2021 and 2022. This was likely due to a policy change in the admissions process, and the Admissions Office may want to review and potentially improve the policy. The role of students' racial ethnicities in their deposit decisions also varied across the years. For instance, Asian and African American students did not show a clear preference for paying deposits in 2020 and 2022, but tended not to pay in certain periods in 2021. In contrast, White students were more likely to pay deposits in 2020 and 2022, but showed no clear preference in 2021. However, Hispanic students were consistently more likely to pay deposits in all three years.

On the other hand, some consistent patterns exist among the three years. Students are more likely to pay deposits in the later periods, because they want to wait for the best admission offers but do not want to miss the deposit deadline. Students are more likely to pay deposits if they show interests to UD, including being willing to be admitted to majors different than the applied majors, and attending campus events such as campus tours and events for prospects and admitted students. Students are less likely to pay deposits, if being from higher income families, living further away from UD, or delaying to review the admission decisions. Although students' socio-economic and demographic backgrounds cannot be changed, we suggest the Admissions Office to survey students who show interests to UD and better target future prospects with similar characteristics.

Chapter 3

STRUCTURAL NEURAL NETWORKS MEET PIECEWISE EXPONENTIAL MODELS FOR INTERPRETABLE COLLEGE DROPOUT PREDICTION

3.1 Introduction

College dropout continues to be a significant concern for higher education institutions (Albreiki et al., 2021; Aulck et al., 2016; Cannistra et al., 2022). Data from the National Center for Education Statistics (NCES) reveals that approximately 40% of first-time, full-time degree-seeking undergraduate students at 4-year degree-granting institutions fail to obtain a bachelor's degree within six years at the same institution. Alarmingly, around 20% of these students drop out within their first year (NCES, 2022). The consequences of dropout are substantial, leading to wasted resources for students, institutions, and society as a whole. Students who discontinue their college education not only waste their time but also the tuition and fees they have paid and the loans they have borrowed, and tended to have lower income and an increased risk of living in poverty (Bouchrika et al., 2023). Concurrently, institutions suffer losses in terms of resources dedicated to these students, as well as potential tuition revenue and future alumni donations. On average, institutions lose approximately $10 million in tuition revenue annually due to attrition (Raisman et al., 2013). Furthermore, both federal and state governments waste their appropriations to institutions and grants to students, with an estimated expenditure of $9 billion between 2003 and 2008 on students who withdrew within their first year (Schneider & Yin, 2010).

Two main approaches have been developed to address the issue of college dropout: theory-driven and data-driven approaches (Cannistra et al., 2022). The theory-driven approach focuses on constructing conceptual models to comprehend the underlying reasons for dropout. It considers students' decisions to discontinue their education as an interplay of various factors, including family background, demographic characteristics, academic performance, social integration, organizational determinants, personal satisfaction, and institutional commitment (Bean & Metzner, 1980; Spady, 1970; Tinto, 1975). By developing these conceptual models and analyzing observed data, specialized recommendations can be derived to mitigate attrition (Bean & Metzner, 1980). In contrast, the data-driven approach emphasizes dropout prediction. Statistical and machine learning models are utilized to forecast students' academic performance and/or identify those who are at risk of dropping out (Aulck et al., 2016; Baker & Yacef, 2009; Heredia et al., 2020; Sultana & Azad, 2017). However, this approach often involves a trade-off between interpretability and predictability. Simpler methods like regression models tend to offer better interpretability but poorer predictive performance, whereas more complex techniques such as neural network models often provide superior predictive accuracy at the cost of reduced interpretability.

Our study aims to integrate the strengths of both approaches, as educational institutions require both explanatory insights and predictive capabilities to develop effective intervention plans for reducing attrition (Wagner et al., 2023). To model college students' dropout risks, we employ a piecewise exponential model (Ameri & Beigi, 2016; Friedman, 1982). This model is well-suited for analyzing longitudinal processes in students' academic careers, which involve time-varying factors and effects associated with dropout risks. While this classic survival analysis model offers interpretability, it may not provide optimal predictive performance. To enhance predictability, we introduce a neural network model into the piecewise survival analysis to capture the hazard, which transitions from a linear combination of variables in traditional survival analysis to a nonlinear function of the variables. However, fully-connected neural network models pose challenges to interpretability due to their black-box nature. To address this dilemma, we employ a structural neural network (Fan et al., 2022; Ranzato et al., 2006; Zhang et al., 2016), where we impose a structure inspired by theoretical frameworks of student attrition onto the neural networks. Specifically, we categorize variables into three groups: academic, economic, and socio-demographic. Variables within each category interact with each other, forming hidden layers that generate a final neuron representing the category. In total, three final neurons are generated. These final neurons are then linearly combined to predict the hazard, which is subsequently converted into dropout probabilities. Consequently, our model not only provides a list of students with a high risk of dropout to a student advising team but also identifies whether students are more likely to dropout due to academic performance (Stinebrickner & Stinebrickner, 2014), financial burden (Cai & Fleischhacker, 2022), or social integration (Stage, 1989).

The primary objective of this chapter is to address three key research questions that will contribute to the design of an effective intervention plan for attrition reduction:

1. Does the utilization of a structural neural network enhance predictive performance compared to traditional survival analysis models that employ linear hazards?

2. Which category of variables impacts students' dropout risks?

3. Do the effects of these variables change as students progress through their academic careers?

3.2 Literature Review and Conceptual Framework

Extensive academic research has been conducted on college student dropout, employing both theory-driven and data-driven approaches. The theory-driven approach aims to establish a theoretical foundation and develop conceptual models to comprehend students' dropout decisions. Tinto (1975) formulated a theoretical model based on Durkheim's suicide theory (Durkheim, 1897) and a cost-benefit analysis, which sought to explain dropout decisions through the interaction between individuals and institutions. The model suggested that factors such as family background, individual characteristics, and pre-college schooling influence individuals' integration within the academic and social systems of colleges, subsequently impacting their commitment to educational goals and institutional commitment. Lower levels of commitment were found to be associated with higher dropout probabilities. Bean (1980) developed a causal model inspired by turnover models in work organizations, positing that the interaction between students' background characteristics and organizational factors affects their satisfaction, institutional commitment, and ultimately dropout probabilities. The model was empirically tested using multiple regression and path analysis, utilizing questionnaires returned by 1,195 new freshmen. Gender-specific recommendations were provided to reduce attrition. Similarly, Spady (1970) developed a sociological model of the college dropout process, drawing from Durkheim's suicide theory (Durkheim, 1897). Spady argued that family background influences students' academic potential and normative congruence, subsequently impacting their grade performance, intellectual development, and friendship support. The interaction of these factors influences social integration, ultimately leading to dropout decisions. These conceptual models have provided theoretical guidance for subsequent empirical studies.

The data-driven approach, as an alternative, emphasizes prediction, specifically the identification of students at high risk of dropout for targeted intervention and remedial programs (Quadri, 2010). Various statistical and machine learning methods have been employed in this approach. Almarabeh (2017) compared five classification methods, including Naïve Bayes, Bayesian network, decision tree with ID3, decision tree with C4.5, and multilayer perceptron neural network, to predict the dropout risks of 225 students using 10 predictors. The Bayesian network model exhibited the best performance across various error measures, such as accuracy, true positive rate, false positive rate, and F-score. Similarly, Sandoval-Palis et al. (2020) compared logistic regression and neural network models to predict the dropout risks of 2,097 students in an engineering university in Ecuador, using four predictors, regime, leveling course type, application grade, and vulnerability index, where regime, leveling course type, and application grade were academic factors, and vulnerability index was derived from 25 socio-economic variables. The neural network models outperformed logistic regression models in terms of accuracy and AUC score. However, despite their high predictive power, neural network models are challenging to interpret due to their black-box nature.

Regression and decision tree models are preferred when the predictive performance is satisfactory, as their interpretability can aid institutions in establishing effective intervention policies. Wagner et al. (2023) compared three explainable methods and two ensemble methods to predict degree dropout in a middle-sized German university. The explainable methods are decision trees, k-nearest neighbors and logistic regression, and the ensemble methods are AdaBoost and Random Forests. Among these, logistic regression was found to exhibit the best overall predictive performance. Nevertheless, the study also highlighted that these models did not equally excel in predicting student subpopulations, particularly concerning gender and specific study programs. Quadri and Kalyankar (2010) proposed a hybrid method for dropout prediction in an Indian institution, combining decision trees to identify relevant predictors, such as parents' income and previous semester's grades, with logistic regression for predicting students' dropout risks. Aulck et al. (2016) utilized regularized logistic regression, random forests, and K-nearest neighbors to predict the attrition of 32,538 students from the University of Washington, finding that regularized logistic regression performed the best. Strong predictors included GPA in math, English, chemistry, and psychology classes. Heredia-Jimenez et al. (2020) employed Random Forest to predict at-risk students across 65 undergraduate programs in a public engineering-oriented university in Ecuador. They obtained reliable predictions by excluding socio-demographics and pre-college entry information, relying solely on academic information as predictors. Ameri et al. (2016) compared the performance of multiple methods, such as logistic regression, adaptive boosting, decision tree, Cox regression, and time-dependent Cox regression, in predicting dropout at Wayne State University. Time-dependent Cox regression was identified as the best method due to its ability to incorporate time-varying predictors, demonstrate superior predictive performance, and predict the timing of dropout events.

Our conceptual framework draws inspiration from Tinto's theoretical model (Tinto, 1975) and aims to capture the longitudinal process of dropout risk accumulation. We posit that students' dropout risks are determined by their goal commitment and institutional commitment, which are influenced by the interplay of academic integration, economic integration, and social integration. Academic integration is shaped by factors such as college cumulative GPA, the number of classes with a grade below D, the number of credits registered, and engagement in multiple majors. Pre-college schooling indicators, such as high school GPA and the number of Advanced Placement (AP) credits, also contribute to academic integration. Economic integration is influenced by variables including expected family contribution (EFC), outstanding balance, and the amount of financial aid received in the form of grants, scholarships, and loans. Social integration encompasses factors such as eligibility for Pell Grants and being a first-generation college student, as well as demographic characteristics including gender and racial ethnicity. Importantly, the dropout risk of a student evolves each semester, reflecting the changing dynamics of the three integrations. These changes stem from time-varying factors like academic performance and financial aid, the evolving processes that generate the integrations, and the dynamic interactions among them.

3.3 Statistical Models

Our statistical model is designed to integrate the theory-driven and data-driven approaches, resulting in an "information-driven" framework (Cannistra et al., 2022). The model combines a structural neural network and a piecewise exponential model to predict dropout probabilities. The piecewise exponential model predicts the dropout probability using a hazard function generated by the structural neural network. The hazard function is a linear combination of three integrations, each of which represents a separate group of variables related to academic, economic, and social factors. For instance, the academic integration is derived solely from academic variables.

3.3.1 Piecewise Exponential Model

We begin with a piecewise exponential model (PEM), a type of discrete event history analysis. An event occurs if a student drops out from the institution they are currently enrolled in by the end of the third year. Otherwise, the student is considered to be "censored" or "survived". For a student i in a semester j, Equation (3.1) defines the logarithm of the hazard function $log(h_i[j])$ to be the sum of baseline hazard $h_0[j]$ and a linear combination of the student attributes $X_{0i}[j]$, where $\beta_0[j]$ represent the coefficients to be estimated for corresponding input variables. The cumulative hazard, calculated as the product of the hazard function and the semester length ($L[j]$), is used to derive the logarithm of the survival function ($S_i[j]$) in Equation (3.2), where $L[j] = 1$ for all semesters. Finally, the student's dropout probability in the semester ($\theta_i[j]$) is calculated as one minus the survival function, as shown in Equation (3.3). While this model allows us to estimate the influence of each input variable on the hazard function and, subsequently, the dropout probability, it is important to note that it assumes a log-linear relationship between the hazard function and the input variables. To account for potential nonlinear relationships between these variables, we will introduce nonlinearity in our subsequent model. Appendix B contains a graphic illustration of this PEM model.

$$log(h_i[j]) = h_0[j] + \beta_0[j]X_{0i}[j] \tag{3.1}$$

$$log(S_i[j]) = -h_i[j]L[j] \tag{3.2}$$

$$\theta_i[j] = 1 - S_i[j] \tag{3.3}$$

3.3.2 Piecewise Exponential Model with Fully-Connected Neural Network

We introduce neural network (NN) to derive the hazard function in a piecewise exponential model. The hazard function is generated from a hidden layer of neurons in this PEM-NN model as shown in Equation (3.5), and the hidden neurons are generated from the input variables as shown in Equation (3.4). The survival function and dropout probability can be calculated in the same way from Equations (3.2) and (3.3). Notably, the inclusion of the neural network introduces nonlinearity into the model, driven by the sigmoid activation function (σ) employed in Equation (3.4). The prediction performance could be improved PEM-NN compared to PEM, because the neural network structure could capture more complex relationships between $X_{0i}[j]$ and $h_i[j]$. However, it's essential to acknowledge that the PEM-NN model comes with a trade-off. For instance, when identifying a group of students at high risk of dropout, discerning the specific reasons behind this elevated risk becomes intricate. Consequently, devising precise and strategic intervention plans based on the model's output becomes a challenging endeavor. Therefore, we need a model to balance interpretation and prediction. For further insights into the PEM-NN model's architecture and visualization, please refer to Addendum B, which provides a graphical representation of this model.

$$X_{1i}[j] = \sigma(\alpha_0[j] + \beta_1[j]X_{0i}[j]) \tag{3.4}$$

$$log(h_i[j]) = h_0[j] + \beta_2 X_{1j} \tag{3.5}$$

3.3.3 Piecewise Exponential Model with Structural Neural Network

We design a structural neural network (SNN) to balance interpretation and prediction. In this PEM-SNN model, the input variables $X_{0i}[j]$ are grouped into three categories, $X_{0i}^{Acad}[j]$ representing academic activities and performance, $X_{0i}^{Econ}[j]$

related to financial aid and financial burden, and $X_{0i}^{Socl}[j]$ representing family background and demographic characteristics. From the input layer, a hidden layer $X_{1i}^{Acad}[j]$ is generated from $X_{0i}^{Acad}[j]$ using Equation (3.6), and an academic neuron $X_{2i}^{Acad}[j]$ is then generated from the hidden layer using Equation (3.7). Similar transformations occur for an economic neuron, $X_{2i}^{Econ}[j]$, and a social neuron, $X_{2i}^{Socl}[j]$, as shown in Equations (3.8) to (3.11). The final three neurons for each integration form the second hidden layer, and the hazard function is the output of the neural network. Equation (3.12) defines the logarithm of the hazard function $h_i[j]$ as the sum of the baseline hazard $h_0[j]$ and a linear combination of the three neurons. The academic integration is represented by $\beta_2^{Acad}[j]X_{2i}^{Acad}[j]$, the economic integration by $\beta_2^{Socl}[j]X_{2i}^{Socl}[j]$, and the social integration by $\beta_2^{Socl}[j]X_{2i}^{Socl}[j]$.

$$X_{1i}^{Acad}[j] = \sigma\left(\alpha_0^{Acad}[j] + \beta_0^{Acad}[j]X_{0i}^{Acad}[j]\right) \tag{3.6}$$

$$X_{2i}^{Acad}[j] = \sigma\left(\alpha_1^{Acad}[j] + \beta_1^{Acad}[j]X_{1i}^{Acad}[j]\right) \tag{3.7}$$

$$X_{1i}^{Econ}[j] = \sigma(\alpha_0^{Econ}[j] + \beta_0^{Econ}[j]X_{0i}^{Econ}[j]) \tag{3.8}$$

$$X_{2i}^{Econ}[j] = \sigma(\alpha_1^{Econ}[j] + \beta_1^{Econ}[j]X_{1i}^{Econ}[j]) \tag{3.9}$$

$$X_{1i}^{Socl}[j] = \sigma\left(\alpha_0^{Socl}[j] + \beta_0^{Socl}[j]X_{0i}^{Socl}[j]\right) \tag{3.10}$$

$$X_{2i}^{Socl}[j] = \sigma\left(\alpha_1^{Socl}[j] + \beta_1^{Socl}[j]X_{1i}^{Socl}[j]\right) \tag{3.11}$$

$$\begin{aligned} log(h_i[j]) = h_0[j] + \beta_2^{Acad}[j]X_{2i}^{Acad}[j] \\ +\beta_2^{Econ}[j]X_{2i}^{Econ}[j] + \beta_2^{Socl}[j]X_{2i}^{Socl}[j] \end{aligned} \tag{3.12}$$

Figure 3.1 illustrates the structure of the hybrid model. The input layer consists of three blocks of variables, which will be described in detail in the Data and Variable section. The academic input block comprises thirteen variables, the economic input block includes six variables, and the social input block contains nine variables. Each input block generates an individual block of neurons in the first hidden layer. The second hidden layer consists of three neurons in total, with each neuron representing one integration: academic, economic, and social. These neurons serve as the intermediate representation of the input variables before generating the hazard function.

The output layer represents the hazard function, which, in turn, generates the survival function and dropout probability using Equations (3.2) and (3.3). It's important to note that the input variables for different integrations do not interact with each other until the process of generating the hazard function. By enforcing this structured design, we ensure that the interactions among input variables are captured at the appropriate stage. If we were to remove this designed structure and allow all input variables to interact with each other from the beginning, the structural neural network would resemble the more traditional fully connected neural network. Conversely, if we were to remove the two hidden layers and directly generate the hazard function as a linear combination of all input variables, this hybrid model would reduce to the simpler piecewise exponential model.

We implement this sparsely connected neural network structure using the Flux package (v0.13.14) (Innes et al., 2018) in Julia, and the code is attached in Appendix C. In a nutshell, the whole network structure is a "chain" in Flux, which is "joined" by six sub-chains, each for one semester. In a chain for a semester, three individual chains are used to transform the three groups of input variables into three neurons, and the three neurons generate the hazard function neuron.

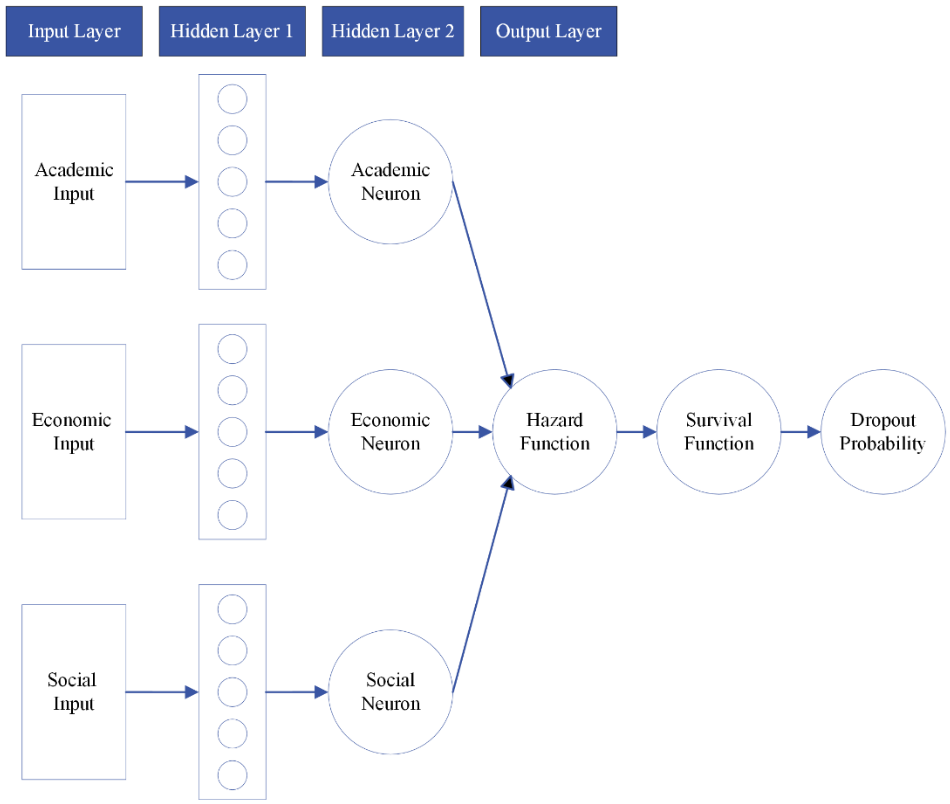

Figure 3.1: A diagram of a hybrid model of a structural neural network and a piecewise exponential model

3.3.4 Piecewise Exponential Model with Structural Neural Network with Interaction Component

We further introduce an interaction neuron between the academic neuron and the economic neuron, because we are interested whether the interaction would play an important role for dropout risk. The process to generate the academic, economic and social neurons are the same with the previous model, and an interaction term $\beta_2[j]X_{2i}^{Acad}[j]X_{2i}^{Econ}[j]$ is added to generate the logarithm of the hazard function $log(h_i[j])$ as shown in Equation (3.13). Addendum B contains a graphic illustration of this model called PEM-SNN-2.

$$log(h_i[j]) = h_0[j] + \beta_2^{Acad}[j]X_{2i}^{Acad}[j] + \beta_2^{Econ}[j]X_{2i}^{Econ}[j]$$
$$+\beta_2^{Socl}[j]X_{2i}^{Socl}[j] + \beta_2[j]X_{2i}^{Acad}[j]X_{2i}^{Econ}[j] \qquad (3.13)$$

3.3.5 Constructing the Loss Function

Suppose a student i was enrolled in a semester j with input variables $X_{0i}[j]$, the loss function in a piecewise exponential model depends on whether the student dropped out by the end of the semester,

$$L_i[j] = \begin{cases} -S_i[j] & \text{(Retained)} \\ -S_i[j]h_i[j] & \text{(Dropout)} \end{cases} \qquad (3.14)$$

We can write the two cases in one equation. Let y_i be the dependent variable, with 1 indicating a student droped out and 0 indicating the student remained enrolled in the next semester. The loss function can be written as

$$L_i[j] = -h_i^{y_i}[j]S_i[j] \qquad (3.15)$$

Therefore, the logarithm of the loss function is

$$logL_i[j] = -y_i[j]log(h_i[j]) - log(S_i[j])$$

According to Equations (2.2) and (2.1), it can be written as

$$logL_i[j] = -y_i[j]log(h_i[j]) + h_i[j]L[j]$$
$$= -y_i[j](h_0[j] + \beta[j]X_{0i}[j]) + exp(h_0[j] + \beta[j]X_{0i}[j])L[j] \qquad (3.16)$$

Where $L[j]$ = 1 for all semesters, regardless of some semesters may last slightly longer than the others. Our dependent variable is imbalanced, because less than 4% of students would drop out in a semester for those who spent their time for at most three years at UD. To address this challenge, we add a hyperparameter w to the loss fuction to make the dropout observations weight more than the others, so the logarithm of loss function becomes

$$logL_i[j] = -w * y_i[j](h_0[j] + \beta[j]X_{0i}[j]) + exp(h_0[j] + \beta[j]X_{0i}[j])L[j] \quad (3.17)$$

For the PEM-NN models, using Equations (3.4) and (3.5) to derive $h_i[j]$, the loss function becomes

$$\begin{aligned} logL_i[j] &= -w * y_i[j]log(h_i[j]) + h_i[j]L[j] \\ &= -w * y_i[j]\left(h_0[j] + \beta_2 X_{1j}\right) + exp\left(h_0[j] + \beta_2 X_{1j}\right)L[j] \\ &= -w * y_i[j]\left(h_0[j] + \beta_2\sigma(\alpha_0[j] + \beta_1[j]X_{0i}[j])\right) \\ &\quad +exp\left(h_0[j] + \beta_2\sigma(\alpha_0[j] + \beta_1[j]X_{0i}[j])\right) \qquad (3.18) \end{aligned}$$

For the PEM-SNN models, using Equation (3.12), the loss function becomes

$$\begin{aligned} logL_i[j] &= -w * y_i[j]log(h_i[j]) + h_i[j]L[j] \\ &= -w * y_i[j](h_0[j] + \beta_2^{Acad}[j]X_{2i}^{Acad}[j] \\ &\quad +\beta_2^{Econ}[j]X_{2i}^{Econ}[j] + \beta_2^{Socl}[j]X_{2i}^{Socl}[j]) \\ &\quad +\exp\,(h_0[j] + \beta_2^{Acad}[j]X_{2i}^{Acad}[j] \\ &\quad +\beta_2^{Econ}[j]X_{2i}^{Econ}[j] + \beta_2^{Socl}[j]X_{2i}^{Socl}[j])L[j] \qquad (3.19) \end{aligned}$$

Where $X_{2i}^{Acad}[j]$, $X_{2i}^{Econ}[j]$ and $X_{2i}^{Socl}[j]$ are derived from Equations (3.6) to (3.11).

For the PEM-SNN-2 models, using Equation (3.13), the loss function becomes

$$\begin{aligned} logL_i[j] &= -w * y_i[j]log(h_i[j]) + h_i[j]L[j] \\ &= -w * y_i[j](h_0[j] + \beta_2^{Acad}[j]X_{2i}^{Acad}[j] \\ &\quad +\beta_2^{Econ}[j]X_{2i}^{Econ}[j] + \beta_2^{Socl}[j]X_{2i}^{Socl}[j]) \\ &\quad +\exp\,(h_0[j] + \beta_2^{Acad}[j]X_{2i}^{Acad}[j] + \beta_2^{Econ}[j]X_{2i}^{Econ}[j] \\ &\quad +\beta_2^{Socl}[j]X_{2i}^{Socl}[j] + \beta_2[j]X_{2i}^{Acad}[j]X_{2i}^{Econ}[j]L[j] \qquad (3.20) \end{aligned}$$

Where $X_{2i}^{Acad}[j]$, $X_{2i}^{Econ}[j]$ and $X_{2i}^{Socl}[j]$ are derived from Equations (3.6) to (3.11).

3.3.6 Model Fitting

To predict the dropout students in a semester, all available data before the semester are used as a training dataset, and the data of the semester are used as a test dataset. The PEM model is trained by the Ipopt solver in the JuMP package (v1.10.0) (Lubin et al., 2023). In the training stage, ten-fold validation and F-score are used to set the values of hyperparameters for the hyperparameter w in the loss function and a threshold probability. The hyperparameter w is set as 1.5 to put more weight on the loss from the dropout observations. After training, hazard function, survival function and dropout probability are calculated using parameter estimates Equations (3.1) to (3.3). The threshold probability is set as 0.15 to determine whether a student would dropout, i.e., a student would drop out if the predicted dropout probability is larger than 0.15; otherwise, the student is predicted to remain enrolled in the next semester.

The hybrid models with neural networks are trained with the Flux package (v0.13.14) (Innes et al., 2018) using Adam optimization algorithm with a learning rate of 0.01, and the exponential decay rates for the first-momentums and second-momentums are 0.9 and 0.999. In addition to the hyperparameter w in the loss function and a threshold probability, the number of neurons is set as 13 using ten-fold validation in the training for the models with fully-connected neural networks. The same number of neurons, i.e., 13, is used for each individual group of neurons derived from their respective variable groups. To ensure the robustness and reliability of the models, 100 runs are conducted to calculate statistical metrics such as precision, recall, and F-score.

3.4 Data and Variables

This study utilizes enrollment and financial aid data from the University of Delaware (UD), a public research university with an undergraduate population of approximately 18,000 students. The dataset includes information from 40,440 undergraduate students who enrolled as first-time full-time students in the fall semesters between 2012 and 2021. The raw data tables were obtained from UD's enterprise data warehouse, and a final dataset for analysis was constructed as a student-semester file, with one row per student and semester.

Students were tracked until the end of their third year or until dropout occurred. Only fall and spring semesters were considered for tracking purposes, and dropout was defined as not being enrolled in the subsequent semester without having graduated. Table 3.1 provides an overview of the headcount flow for the ten cohorts across different semesters. The smallest cohort, Fall 2020, consisted of 3,711 first-time full-time students, while the largest cohort, Fall 2017, comprised 4,258 students. The headcount of each cohort gradually decreased over the semesters due to dropout. Blank cells in the table indicate that the corresponding cohort had not reached that particular semester.

Table 3.1: Cohort headcount flow by semester

Cohort\Semester Sequence	1st	2nd	3rd	4th	5th	6th
Fall 2012	3,806	3,677	3,490	3,398	3,308	3,284
Fall 2013	3,794	3,660	3,438	3,340	3,244	3,197
Fall 2014	4,168	4,035	3,835	3,735	3,642	3,569
Fall 2015	4,092	3,944	3,725	3,644	3,538	3,497
Fall 2016	3,943	3,807	3,576	3,479	3,385	3,297
Fall 2017	4,285	4,116	3,827	3,708	3,582	3,481
Fall 2018	4,242	4,091	3,849	3,732	3,658	3,609

Fall 2019	4,136	3,934	3,705	3,621	3,554	3,484
Fall 2020	3,711	3,573	3,389	3,267	3,171	3,117
Fall 2021	4,263	4,079	3,852	3,775		

Table 3.2 presents the headcounts for each semester between Fall 2012 and Spring 2022, organized by cohort. Each semester featured up to three cohorts: freshman, sophomore, and junior. Senior cohorts were excluded from the analysis as they had already completed their third year or exceeded the sixth semester. The initial semesters do not have three cohorts due to the absence of data before the Fall 2012 cohort. To predict at-risk students in a given semester, the training dataset included all available data up until that semester, while the data specific to the semester served as the test dataset. Initially, the focus was on predicting at-risk students among the 10,830 students enrolled in Spring 2022, using all data preceding Spring 2022 for training. The process was then repeated for Spring 2021 and Spring 2020 to predict at-risk students in those semesters.

Table 3.2: Headcounts between Fall 2012 and Spring 2023 by cohort

Semester \ Cohort	Freshman Cohort	Sophomore Cohort	Junior Cohort	Total
Fall 2012	3,806			3,806
Spring 2013	3,677			3,677
Fall 2013	3,794	3,490		7,284
Spring 2014	3,660	3,398		7,058
Fall 2014	4,168	3,438	3,308	10,914
Spring 2015	4,035	3,340	3,284	10,659
Fall 2015	4,092	3,835	3,244	11,171
Spring 2016	3,944	3,735	3,197	10,876
Fall 2016	3,943	3,725	3,642	11,310
Spring 2017	3,807	3,644	3,569	11,020
Fall 2017	4,285	3,576	3,538	11,399
Spring 2018	4,116	3,479	3,497	11,092

Fall 2018	4,242	3,827	3,385	11,454
Spring 2019	4,091	3,708	3,297	11,096
Fall 2019	4,136	3,849	3,582	11,567
Spring 2020	3,934	3,732	3,481	11,147
Fall 2020	3,711	3,705	3,658	11,074
Spring 2021	3,573	3,621	3,609	10,803
Fall 2021	4,263	3,389	3,554	11,206
Spring 2022	4,079	3,267	3,484	10,830

Table 3.3 provides a description of the dependent and independent variables, along with their means (and standard deviations for numeric variables) for each semester. The dependent variable indicates whether a student is enrolled in the next semester, with a value of 1 representing dropout and 0 representing enrollment. The average dropout rate increased from the first semester (0.033) to the second semester (0.047) and then gradually decreased in the subsequent four semesters: 0.025, 0.023, 0.013, and 0.01 from the third to the sixth semester, respectively.

There are thirteen academic variables, consisting of four binary variables and nine numeric variables. In the first fall semester, all students were full-time students, and the majority remained full-time throughout their academic journey. During the first semester, 16.9% of students were in the University Study program, which allowed them to explore and choose majors later. As students progressed to their sophomore and junior years, the percentage of students in the University Study program gradually decreased to 0.9% by the sixth semester. In the first semester, no students had a double major, but the percentage of students with a double major increased to 14.4% by the sixth semester.

Due to the data collection process on the census day of each semester (two weeks after the semester starts), students' academic standing status was not available in the first semester. There were 6.4% of students in probation status in the 2nd semester, and the proportion gradually decreased to 1.3%. Similarly, these variables are not available in the 1st semester, College GPA, DFW Count, Listener Count, Academic Standing, and Total Credit were not recorded for the first semester. In addition, the Total Credit, Total Minor Credit, Total Transfer Credit, and Total AP Credit variables were standardized by the typical 120 credits required for graduation, and the Registered Credit variable was standardized based on the typical total of 15 credits taken in one semester. For example, total credit rate being 0.25 means that a student has 30 (0.25x120) credits in total as of the census day of a semester, and registered credit rate being 0.8 means that a student registered 12 ((0.8x15)) credits in a semester.

From the second semester to the sixth semester, the average cumulative GPA increased from 3.103 to 3.229, the average DFW class count decreased from 0.371 to 0.249, the average Listener class count decreased from 0.062 to 0.034, and the average total credit rate increased from 0.18 to 0.716. From the first semester to the sixth semester, the average registered credit rate decreased from 1.02 to 0.991, the average total minor credit rate increased from 0.001 to 0.048, the average total transfer credit rate increased from 0.017 to 0.026, and the average total AP credit rate increased from 0.034 to 0.04. The average high school GPA varied slightly between 3.758 and 3.766 due to changes in the student population each semester.

All economic variables are numeric and standardized by the cost of attendance (COA). For instance, a grant/scholarship rate of 0.2 indicates that 20% of the COA is covered by grants and scholarships. Among the six semesters, the average EFC rate varied between 0.819 and 1.005, the average grant/scholarship rate varied between 0.165 and 0.228, the average PLUS parent loan rate varied between 0.036 and 0.053, the average student loan rate varied between 0.128 and 0.147, the average work-study aid rate varied between 0.003 and 0.006, and the balance varied between -0.012 and -0.016. The balance for the first semester is not available as of the census day.

All social variables, except for Age, are binary variables. From the first semester to the sixth semester, the average age increased from 18.023 to 20.442. The proportion of Pell Grant recipients decreased from 15.1% to 12.6%, the proportion of first-generation college students decreased from 12.8% to 11.8%, the proportion of male students varied between 40.0% and 40.5%, the proportion of Delawarean students increased from 31.7% to 32.7%, the proportion of African American students decreased from 5.2% to 4.6%, the proportion of Asian students varied between 5.1% and 5.3%, the proportion of Hispanic students decreased from 9% to 8.2%, and the proportion of White students increased from 72.8% to 74.4%.

Table 3.3: Description and mean (standard deviation) of input and output variables by semester; standard deviation is only calculated for numeric variables.

	Description	1st	2nd	3rd	4th	5th	6th
Output Variable							
Dropout	Whether a student not enrolled next semester	0.033	0.047	0.025	0.023	0.013	0.01
Academic Variable							
Full-time status	Whether a full-time student	1	0.995	0.994	0.993	0.992	0.988
University Study	Whether in University Study (US) program	0.169	0.15	0.099	0.052	0.019	0.009
Double major	Whether a student has double major	0	0.036	0.081	0.116	0.136	0.144
College GPA	Cumulative college GPA	N/A	3.103 (0.692)	3.143 (0.586)	3.18 (0.535)	3.197 (0.509)	3.229 (0.488)
DFW count	Count of classes with final grade being D (D+, D, D-), F or W (withdraw) in last	N/A	0.371 (0.849)	0.311 (0.758)	0.326 (0.775)	0.265 (0.709)	0.249 (0.696)

	semester						
Listener count	Count of audit (Listener) classes in last semester	N/A	0.062 (0.25)	0.059 (0.255)	0.045 (0.239)	0.042 (0.22)	0.034 (0.204)
Academic standing	Whether a student is in probation status	N/A	0.064	0.039	0.028	0.017	0.013
Total credit	Number of total credits over 120	N/A	0.18 (0.083)	0.308 (0.092)	0.447 (0.103)	0.574 (0.113)	0.716 (0.121)
Registered credit	Number of credits taken in the current semester over 15	1.02 (0.077)	1.014 (0.103)	1.014 (0.107)	1.003 (0.114)	1.009 (0.123)	0.991 (0.139)
Total minor credit	Number of credits earned from winter/summer terms over 120	0.001 (0.005)	0.008 (0.016)	0.013 (0.022)	0.026 (0.032)	0.032 (0.037)	0.048 (0.045)
Total transfer credit	Number of total transfer credits over 120	0.017 (0.045)	0.017 (0.045)	0.017 (0.044)	0.022 (0.047)	0.022 (0.046)	0.026 (0.049)
Total AP credit	Number of total AP credits over 120	0.034 (0.06)	0.035 (0.061)	0.037 (0.062)	0.038 (0.063)	0.039 (0.064)	0.04 (0.064)
HS GPA	High school GPA	3.759 (1.265)	3.766 (1.285)	3.758 (1.373)	3.763 (1.389)	3.755 (1.467)	3.755 (1.382)
Economic Variable							
EFC	Expected family contribution (EFC) over cost of attendance (COA)	0.997 (1.488)	1.005 (1.487)	0.849 (1.137)	0.855 (1.147)	0.819 (1.052)	0.824 (1.065)
Grant	Total grant/scholarship obtained over COA	0.22 (0.211)	0.228 (0.217)	0.185 (0.199)	0.194 (0.206)	0.165 (0.19)	0.175 (0.198)
PLUS loan	Loan borrowed by parent over COA	0.047 (0.165)	0.053 (0.174)	0.04 (0.155)	0.049 (0.17)	0.036 (0.145)	0.045 (0.161)
Student loan	Loan borrowed by student over COA	0.128 (0.182)	0.135 (0.198)	0.129 (0.183)	0.142 (0.206)	0.131 (0.181)	0.147 (0.205)
Work study	Financial aid in work study over COA	0.006 (0.016)	0.005 (0.015)	0.003 (0.013)	0.003 (0.013)	0.003 (0.012)	0.003 (0.013)
Balance	Outstanding balance over COA	N/A	-0.013 (0.054)	-0.012 (0.051)	-0.014 (0.055)	-0.011 (0.049)	-0.016 (0.059)
Social Variable							
Pell	Whether a student received federal Pell grant	0.151	0.147	0.134	0.132	0.126	0.126
First Generation	Whether a first-generation college student	0.128	0.124	0.121	0.119	0.119	0.118
Male	Whether a male student	0.405	0.404	0.403	0.4	0.404	0.402
Residency	Whether a Delawarean student	0.317	0.317	0.324	0.323	0.327	0.327

Age	Age in the current semester	18.023 (0.554)	18.449 (0.651)	19.021 (0.516)	19.446 (0.572)	20.018 (0.518)	20.442 (0.571)
African American	Whether an African American student	0.052	0.051	0.049	0.048	0.047	0.046
Asian	Whether an Asian student	0.052	0.052	0.051	0.051	0.052	0.053
Hispanic	Whether a Hispanic student	0.09	0.089	0.086	0.085	0.082	0.082
White	Whether a White student	0.728	0.73	0.735	0.739	0.743	0.744

3.5 Results and Discussion

We conducted a comparison of four models for predicting dropout students in Spring 2020, Spring 2021, and Spring 2022. The models compared were the piecewise exponential model with a linear hazard function (PEM-Linear), the piecewise exponential model with a fully-connected neural network for the hazard function (PEM-NN), the piecewise exponential model with a structural neural network for the hazard function (PEM-SNN), and the piecewise exponential model with a structural neural network with interaction for the hazard function (PEM-SNN-2), where the interaction occurs between the academic and economic integrations. For each spring semester, the PEM-Linear model was executed once using the JuMP package (v1.10.0) (Lubin et al. 2023), while the other models were run 100 times using the Flux package (v0.13.14) (Innes et al. 2018). This distinction was necessary because a global optimization could only be found for the PEM-Linear model.

Table 3.4 presents a comparison of the four models in terms of their average precision, recall, and F-score, along with their standard deviations. The model with the highest average F-score is highlighted in bold for each spring semester. A student is predicted to have a high dropout risk if their predicted dropout probability is greater than 0.15. Precision indicates the proportion of students who actually dropped out and are predicted to have a high risk out of all students predicted to have a high risk. A higher precision implies a higher proportion of interventions can be directed towards students who truly require help, thus making any intervention program more efficient. Recall reflects the proportion of students who actually dropped out and are predicted to have a high risk out of all students who actually dropped out. Higher recall indicates that a larger proportion of dropout students would be covered by the interventions, making the intervention program more effective. The F-score is the harmonic mean of precision and recall, measuring both the efficiency and efficacy of the intervention program.

Our discussion on model performance focuses on their F-scores, with PEM-Linear serving as the baseline performance among the four models. We use either the z-test or Kolmogorov-Smirnov (KS) test in the HypothesisTests package (v0.10.11) to determine whether the F-scores of two models are similar or not. For Spring 2020, PEM-NN shows similar F-scores (average = 0.172) to PEM-Linear (0.173), while PEM-SNN and PEM-SNN-2 demonstrate higher F-scores with averages of 0.179 and 0.176, respectively. In Spring 2021, PEM-NN exhibits higher F-scores (average = 0.214) compared to PEM-Linear (0.209), while PEM-SNN and PEM-SNN-2 achieve even higher F-scores with averages of 0.229 and 0.232, respectively. In Spring 2022, PEM-NN, PEM-SNN, and PEM-SNN-2 demonstrate similar F-scores to PEM-Linear (0.251), with averages of 0.251, 0.250, and 0.249, respectively. Therefore, PEM-SNN and PEM-SNN-2 exhibit similar or higher predictive performance than the other two models. As PEM-SNN and PEM-SNN-2 exhibit similar performance, but PEM-SNN is easier to interpret than PEM-SNN-2, we select PEM-SNN as the best model and focus on its interpretation in the following section.

Table 3.4: Performance comparison of four models in five semesters in terms of average (and standard deviation) of precision, recall and F-score

		Spring 2020	Spring 2021	Spring 2022
PEM-Linear	Precision	0.143	0.303	0.227
	Recall	0.221	0.160	0.280
	F-Score	0.173	0.209	**0.251**
PEM-NN	Precision	0.132 (0.011)	0.244 (0.025)	0.221 (0.01)
	Recall	0.246 (0.024)	0.193 (0.018)	0.29 (0.011)
	F-Score	0.172 (0.015)	0.214 (0.013)	**0.251 (0.006)**
PEM-SNN	Precision	0.137 (0.007)	0.224 (0.013)	0.213 (0.011)
	Recall	0.258 (0.017)	0.236 (0.026)	0.304 (0.022)
	F-Score	**0.179 (0.009)**	0.229 (0.014)	0.250 (0.011)
PEM-SNN-2	Precision	0.135 (0.007)	0.228 (0.015)	0.212 (0.011)
	Recall	0.253 (0.015)	0.238 (0.025)	0.303 (0.022)
	F-Score	0.176 (0.008)	**0.232 (0.014)**	0.249 (0.012)

We begin the interpretation of PEM-SNN for predicting dropout students in Spring 2020 using a model with an F-score close to the average from the 100 runs. In Spring 2020, there are three cohorts: the Fall 2019 cohort in their second semester, the Fall 2018 cohort in their fourth semester, and the Fall 2017 cohort in their sixth semester. Figure 3.2 presents histograms of the predicted logarithm of hazard (log(h)) and the three integrations, where log(h) is the sum of the intercept (not shown in the figure) and the three integrations as per Equation (3.7).

The first row of Figure 3.2 displays the distributions of log(h) for retained and dropout students. Log(h) is positively related to the dropout probability, i.e., the higher the log(h) is, the higher the dropout probability is, and vice versa. Among retained students, the distribution for the Fall 2019 cohort exhibits a relatively symmetric bell shape centered around -3, the distribution for the Fall 2018 cohort is more concentrated around -4.8 with a long right tail, and the distribution for the Fall 2017 cohort is mostly concentrated near -5.3. According to Equations (3.2) and (3.3), a smaller log(h) corresponds to a lower dropout probability. Therefore, for retained students, their dropout probabilities gradually decrease as they progress in their academic careers, and it becomes increasingly certain that they will not dropout. The distributions of log(h) for dropout students differ from those of retained students. The distribution for the Fall 2019 cohort becomes flatter and shifts towards larger log(h) values, while the distributions for the Fall 2018 and Fall 2017 cohorts are even flatter and spread between -6 and 0. This pattern aligns with the expectation that dropout students would have higher log(h) values and thus higher predicted dropout probabilities.

The subsequent rows of the figure display histograms of the three integrations. The academic integration contributes the most to distinguishing between retained and dropout students. Similar to log(h), the distributions of the academic integrations for dropout students are flatter and shift towards higher values compared to those of retained students. However, the distributions of the other two integrations show similar patterns between dropout and retained students. Additionally, the magnitudes of the other two integrations can be too small to make a meaningful contribution to log(h), including the social integration for the Fall 2018 cohort and the economic and social integrations for the Fall 2017 cohort. Based on these findings, if there were an intervention program in Spring 2020, it should focus on students' academic performance and activities to reduce attrition.

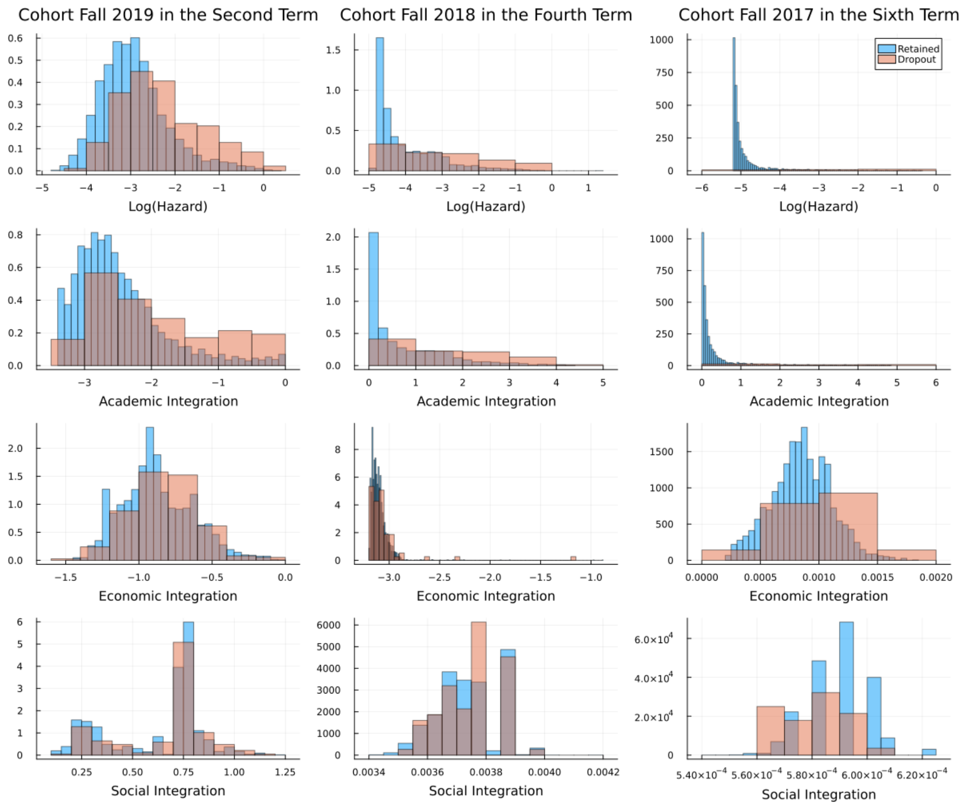

Figure 3.2: Histograms of log(hazard), academic integration, economic integration, and social integration for cohort Fall 2019, cohort Fall 2018 and cohort Fall 2017 in Spring 2020

Figure 3.3 presents histograms of log(h) and the three integrations for cohort Fall 2020, cohort Fall 2019, and cohort Fall 2018 in Spring 2021. The log(h) exhibits a similar pattern to Spring 2020, where retained students' log(h) values move towards smaller values as they progress in their academic careers, with the log(h) being highly concentrated for students in their sixth term. On the other hand, dropout students' log(h) values are more dispersed towards higher values, indicating a higher predicted dropout risk. Similarly to Spring 2020, the academic integrations display a similar magnitude and pattern as log(h), contributing to the larger hazard among dropout students compared to retained students. The economic integrations' magnitudes gradually decrease as students progress academically, suggesting their diminishing influence on dropout risk. They do not contribute much to differentiate between retained and dropout students, as their distributions are similar between the two groups. The social integrations exhibit relatively stable magnitudes but display similar patterns for both retained and dropout students, indicating limited contribution to differentiating between the two groups.

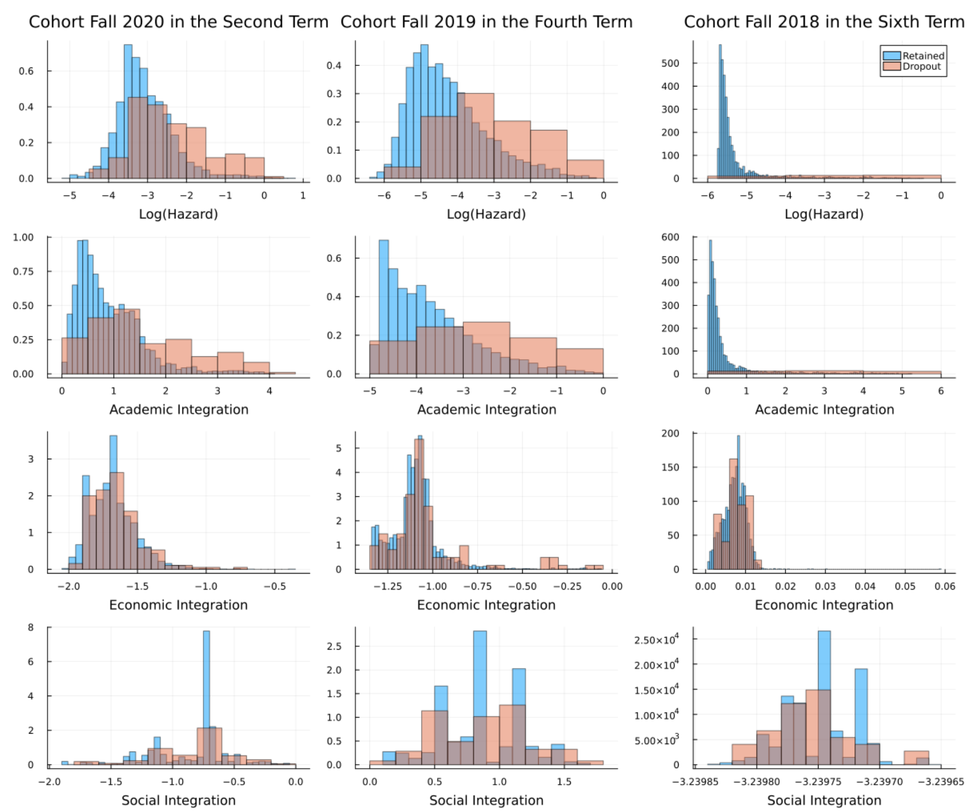

Figure 3.3: Histograms of log(hazard), academic integration, economic integration, and social integration for cohort Fall 2020, cohort Fall 2019 and cohort Fall 2018 in Spring 2021

Figure 3.4 displays the histograms for cohort Fall 2021, cohort Fall 2020, and cohort Fall 2019 in Spring 2022. Similar to the findings in Spring 2020 and Spring 2021, the log(h) distributions of dropout students have a flatter shape compared to retained students, particularly for the sophomore and junior cohorts. Furthermore, the distribution of log(h) shifts towards higher values for the dropout students, indicating a higher predicted dropout probability. The distributions of academic integrations also exhibit distinct patterns between retained and dropout students, with the latter group tending to have higher academic integrations. In contrast, the economic and social integrations demonstrate similar distributions across both student groups. Consequently, our consistent findings across Spring 2020, Spring 2021, and Spring 2022 underscore the crucial role of academic integration in differentiating between dropout and retained students.

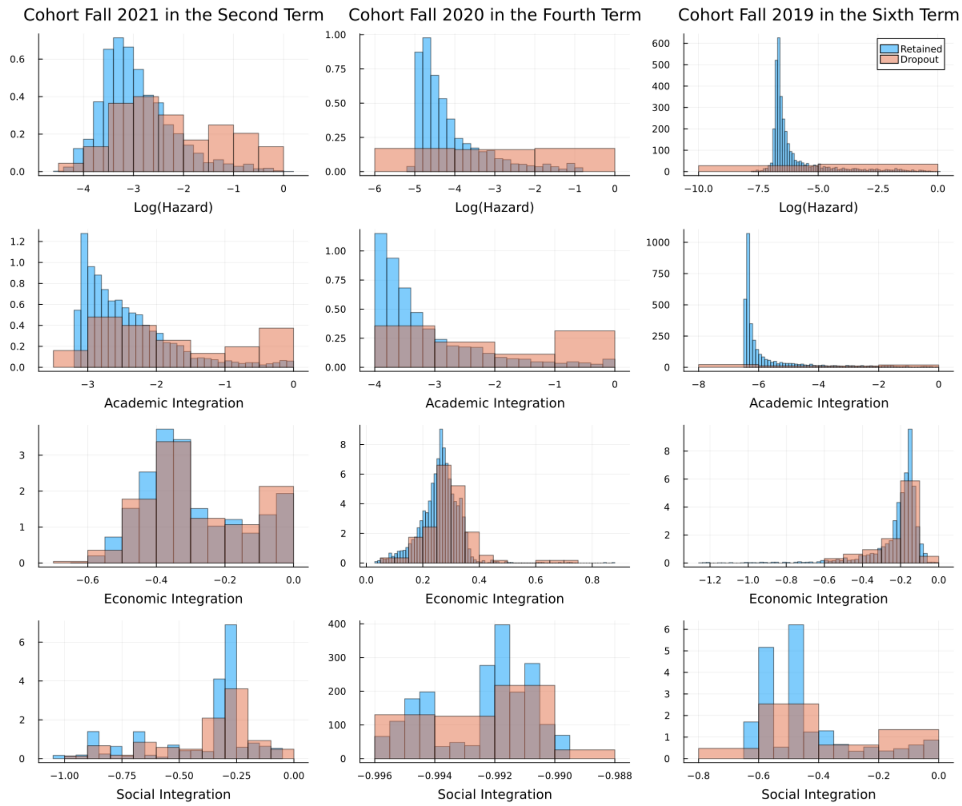

Figure 3.4: Histograms of log(hazard), academic integration, economic integration, and social integration for cohort Fall 2021, Fall 2020 and Fall 2019 in Spring 2022

3.6 Conclusion

We have developed a hybrid model that combines a structural neural network and a piecewise exponential model, aiming to assist in the design of an intervention plan to mitigate student attrition in colleges. This hybrid model not only enables us to predict a list of students with a high risk of dropout but also provides potential explanations for the elevated risks. We apply the model to predict dropout students during the Spring semesters of 2020, 2021, and 2022 at the University of Delaware, and compare its performance with two other models. Whether incorporating an interaction term between academic and economic integrations or not, the hybrid model consistently demonstrates similar or superior predictive performance compared to a piecewise exponential model with linear hazard and another hybrid model featuring a more traditional fully-connected neural network.

Among the three integrations, the academic integration proves to be the most influential in distinguishing between dropout and retained students, as indicated by the results obtained from the PEM-SNN models, consistent with prior research (Aulck et al., 2019; Berens et al. 2019). Dropout students are expected to exhibit higher predicted dropout probabilities compared to their retained counterparts, with the dropout probabilities being determined by the logarithm of the estimated hazards. While the hazard distributions of retained students shift towards lower values, the distributions of dropout students are more dispersed and encompass the higher hazard range, confirming that dropout students indeed tend to have higher hazards than retained students. The distributions of the academic integrations follow a similar pattern to the hazards, displaying noticeable differences between dropout and retained students, while the distributions of the economic and social integrations appear similar between the two groups of students. This consistent finding persists across different stages of students' academic careers, irrespective of whether it is the second, fourth, or sixth semester. Based on these findings, we recommend that intervention programs focus on closely examining the academic performance and activities of at-risk students.

Chapter 4

THE EFFECT OF LOAN DEBT ON GRADUATION BY DEPARTMENT: A BAYESIAN HIERARCHICAL APPROACH

4.1 Introduction

Increasing costs for higher education leads to college students accumulating substantial amounts of loan debt (College Board, 2020; Hemelt et al., 2018; Middaugh, Graham, and Shahid, 2003). Student loans are the second largest category of debt that Americans take on - exceeding $1.6 trillion nationally in the fourth quarter of 2019 (Johnson, 2019; USAFacts, 2019). The financial burden of loan debt affects the students’ persistence, graduation, and lives after graduation (Herzog, 2018; Noopila and Williams Pichon, 2020; Patel, 2020). Even some 2020 presidential candidates, notably Sen. Bernie Sanders and Sen. Elizabeth Warren, proposed costly student loan debt forgiveness plans arguing that students struggle with basic expenses due to repaying student loans (Johnson, 2019; Patel, 2020).

Unlike the U.S. government, educational institutions do not have the resources to fully cancel students' loan debt. However, colleges and universities can certainly design financial aid policies that ensure better student outcomes. This paper will investigate the relationship between student loan debt and six-year graduation rate. Graduation rate is a standard measure of student outcome, e.g., Integrated Postsecondary Education Data System (IPEDS) reports six-year graduation rate for first-time, full-time undergraduate students seeking a bachelor's degree at four-year, degree-granting institutions that participate in Title IV financial aid programs.

Previous studies report mixed findings on the effect of loan debt on graduation. Loan debt has been seen as beneficial (Bowen, Chingos, and McPherson, 2009), detrimental (Chen and Hossler, 2017; Davidson and Holbrook, 2014; Franke, 2019; Jones-White et al., 2014), and non-significant (Dowd and Coury, 2006; Gross, Torres, and Zerquera, 2013). Given that loan debt is confounded with other factors such as student demographic characteristics, socioeconomic status, and academic performance we might expect these mixed results as different contexts and different controls will certainly alter findings.

Previously considered control factors include demographic characteristics such as gender and age, standardized tests such as SAT/ACT scores and number of AP credits, academic performance such as number of courses with C or D grade, socioeconomic status such as adjusted gross income (AGI), and institutional type if a study involves multiple institutions (Gross, Torres, and Zerquera, 2013; Jones-White et al., 2014; Noopila and Williams Pichon, 2020). In the literature, most studies focus on particularly relevant subsets of potential control factors. For instance, Gross et al. and Zhan et al. focused on the effect of loan debt by racial and ethnic group (Gross, Torres, and Zerquera, 2013; Zhan, Xiang, and Elliott III, 2018). The earlier study found that loan debt did not directly affect Latino/a graduation. The latter study showed that minority students had higher levels of tolerance for loan debt when compared with their Caucasian counterparts. Another study focused on the effect of loan debt by students' economic background and institutional type (public/private) (Dwyer et al., 2012) and found that the graduation rates of students in private institutions and those from upper-income families attending public institutions were less affected from loan debt.

This study builds on the previous literature by being the first to include a student's department when considering student loan effects on graduation rate. Similar to the observation that financial aid has different effects in different racial and ethnic groups or socioeconomic groups (Kim, 2012), the financial burden from loan debt will likely vary with income prospects (Witteveen and Attewell, 2019). For example, the average median income is less than $21,000 for Drama/Theatre Arts and Stagecraft majors among public institutions, but it is more than $68,000 for Computer Engineering majors (U.S. Department of Education, 2019). According to the human capital model developed by Gary Becker (Becker, 1964), a student tends to drop out from school and enter the labor market when the marginal cost of education is higher than the expected benefit of education (Long, 2007). This is an undesirable outcome for the institutions because graduation rates are prominent in both government funding criteria and popular-press school-ranking criteria (Lin, 2020).

As will be shown, our findings suggest that institutions should consider students' departments when distributing institutional financial aid resources. Universities can pay special attention to students in departments where graduation rates seem to suffer under the burden of loan debt. One response might be to offer more institutional scholarships to students in those departments, better aligning debt burden with future outcomes. Our study also suggests that departmental variation in induced debt burden varies with academic level (e.g. freshman, sophomore, etc.), and hence, financial aid policy should consider differences between first-year students and more advanced students in the department.

To estimate the effect of control factors on graduation, previous studies typically use different variants of logit models. Dowd and Coury used a standard dichotomous logit model to estimate the effects of control factors (Dowd and Coury, 2006; Gross, Hossler, and Ziskin, 2007; Noopila and Williams Pichon, 2020). Jones-White et al. used a multinomial logit model for their outcome variable which had three levels: graduate from the initial university, graduate from a transfer university, or not graduate (Jones-White et al., 2014). Moreover, previous studies utilized event history models to investigate the effect of loan debt, to include time-varying covariates such as cumulative loan debt (Chen and Hossler, 2017; Gross, Torres, and Zerquera, 2013). However, event history models typically assume proportional hazard, i.e., the effects are time-independent for the time-varying covariates, while this study assumes the effects of the time-varying covariates are also time-varying.

Our study also contains a model within the logit-family of models. Specifically, we use a Bayesian hierarchical logit model to estimate loan debt effect on graduation by department. Since some departments have few students enrolled, we chose Bayesian estimation to overcome the challenges presented by small data. While classical regression methods of previous studies are susceptible to over-fitting due to small sample sizes, the Bayesian hierarchical approach uses partial pooling to overcome this difficulty (McElreath, 2015). The inputs of the models are students' information at the end of a spring semester, such as students' departments, cumulative loan debt, demographic characteristics, academic preparation, academic performance, and financial background. The output of the models is whether students graduate within six years. Since this study intends to propose different financial aid policies for enrolled students with different academic levels, separate models are built for first-year students, second-year students, third-year students, fourth-year students and fifth-year students, respectively.

In summary, this chapter investigates the following two research questions in order to encourage institutions to consider students' departments in their financial aid policies:

1. Does loan debt have the same effect on six-year graduation for enrolled students in different departments in the same semester?

2. Does the effect of loan debt in a department change over time?

4.2 Prior Research

This section starts with a broad discussion of prior research on the effects of financial aid on student choices, and then narrows the discussion to look at the effect of loan debt on graduation. Two key aspects of prior research inspire this study. First, loan debt effect on college graduation is still inconclusive. Second, the effect of loan debt on college graduation can be different in different student groups. Hence, designing differential financial aid policies by department might prove an effective way to enhance students' outcomes and to deploy financial resources efficiently.

4.2.1 Financial Aid on Student Choices

The prior research on financial aid focused on federal and state financial aid policies, and the effect of those policies on sequences of student choices: whether to attend college, which college to attend, whether to persist, and which major to study (Dynarski and Scott-Clayton, 2013; St. John, 1991). After reviewing the studies from 1970s to 2010s, St. John, Dynarski and Scott-Clayton summarized the following findings. First, increased financial aid, including loans, promotes the decision to attend college with larger effect sizes seen in middle-income student populations as compared to low-income student populations (Leslie and Brinkman, 1988; St John, 1990). Second, middle-income students prefer less expensive institutions if they need to finance the cost of attendance with loans (McPherson and Schapiro, 2010; van der Klaauw, 2002). Third, financial aid promotes persistence when studied nationally, but this retention-type of effect is not consistently seen in institutional research (Carroll, 1987; Leslie and Brinkman, 1988; Moline, 1987) or for late-year retention (St. John, 1989). Fourth, loan debt does not affect the choice of major (St. John and Noell, 1987). Fifth, program complexity decreases the effectiveness of financial aid (Bettinger et al., 2012). At the end of his review, St. John calls for more research to help the design of both institutional financial aid policies and to refine both financing and enrollment management strategies.

4.2.2 Effect of Loan Debt on graduation

Findings of the effect of loan debt on graduation have been inconclusive. Many previous studies found loan debt to be a detrimental effect on college graduation. Focusing on first-time students in their first term, Jones-White et al. (2014) concluded that loan aid encouraged college students to quit entirely or transfer to another institution. Specifically, the relative risk of not graduating from the first-entry institution increased by about 7.51%, with the increase of $1,000 in loan debt. While also studying first-time first-year students, Gross et al. (2013) focused on the effect of loan debt on graduation by race/ethnicity in Indiana's public four-year institutions. They found a negative relationship between loan debt and graduation for African American or Black students. Specifically, the odds of graduation of an African American or Black student decreased by 0.2% with an increase of $100 in loans, compared to White students. In studying the effect of financial aid on graduation of non-traditional students, Chen and Hossler (2017) found federal unsubsidized student loans had a detrimental effect on timely graduation. Specifically, the probability of graduating within six years decreased by 1%, with the increase of $1,000 in unsubsidized loan debt.

Loan debt was also found to have an indirect effect on college graduation. Focusing on first-semester students in community colleges, Dowd and Coury (2006) observed no significant effect from subsidized loans on graduation within five years. When investigating loan effect along with parental income and institution type, Dwyer, McCloud, and Hodson (2012) found the effect of loan debt was not significant for students in private schools. Conversely, for African American or Black students, Gross et al. (2013) found that loan debt did not have a direct effect on graduation for Latino/a students.

Surprisingly, a few studies report positive effects of loan debt. For example, Bowen et al. (2009) suggested that federal loans have a positive effect on graduation. However, the conjecture was based on the positive effect of loans on college attendance and was not supported by an empirical study.

Some studies found the relationship between loan debt and college graduation was nonlinear. Dwyer et al. (2012) discovered the effect was positive if borrowing was less than about $10,000 but was negative if beyond $10,000 for public university students from modest economic backgrounds. Focusing on the difference by race and ethnicity, Zhan et al. (2018) confirmed the nonlinear relationship of the effect of loan debt on graduation. They claimed loan debt had a positive effect on graduation until borrowing more than $18,452, $20,990, $23,971 for White, Black, and Hispanic students, respectively.

Another branch of the literature investigates the effects of loan debt on graduation for different student groups. Race/ethnicity is frequently used to group students when investigating these effects. The goal of these lines of inquiry is to find whether financial aid helps to improve equity in education. As mentioned earlier, although they found that loan debt did not have a direct effect on graduation for Latino/a students, Gross et al. (2013) found a negative relationship between loan debt and graduation for African American or Black students. In Zhan et al.'s (2018) study, they found a consistent nonlinear relationship between loan debt and graduation among students with different races/ethnicities but also found that minority students had a higher tolerance for loan debt when compared with White students. In addition to race/ethnicity, institutional type is another popular factor used to group students. Institutional type is an important factor related to graduation rate (Gross, Torres, and Zerquera, 2013; Hussar et al., 2020), loan debt amount (NCES, 2020), and loan default (Hillman, 2014). Dwyer et al. (2012) found students with low and medium income at public universities were negatively affected by loan debt, while loan debt did not affect graduation for students at private universities. Due to data availability, this study includes data from a single public institution and thus, does not include institutional type as a control variable.

Although students' major/department of study has been considered in the research related to loan debt amount (Burns and Webber, 2019; George-Jackson, Rincon, and Martinez, 2012; Harrast, 2004) and loan default (Gross et al., 2009; Hillman, 2014), no previous research studied the effect of loan debt by major/department on college graduation. We expect more studies will include this factor in the future. For example, in May 2019 and December 2019, the U.S.

Department of Education has published preliminary and official files on loan debt and income by field of study (U.S. Department of Education, 2019). The files reveal that the accumulated loan debt and earnings after graduation are significantly different in different fields of study. With the different expected income after graduation, students may feel different financial stress from loan debt. Students with excessive financial stress may choose to drop out (Fossey and Bateman, 1998). An institution could design a differential financial aid policy by field of study (Luna-Torres et al., 2018), according to the different effects of loan debt on graduation in different fields. However, since an institution is typically organized by department instead of field of study, it may be more practical to design the policy by department.

Additionally, our study overcomes some (not all) of the generalizability concerns of previous studies by including multiple cohorts of students and studying them across multiple academic levels. Jones-White et al. (2014) expressed concern that their results were based on only first-term financial aid information from a single cohort of students. Dowd and Coury obtained the data from Beginning Postsecondary Students Longitudinal Study (BPS 90/94). They expressed concern that the financial information of students might change in subsequent years, but only the first-year financial information was available in BPS 90/94 (Dowd and Coury, 2006). This study includes five years of financial aid information from students admitted in three different fall semesters. In addition, we do not assume loan debt effect is the same over years, and the effects in different years are investigated in separate models.

4.3 Conceptual Framework

This study adapts the Student Adjustment Model to understand students' decisions regarding college graduation (Cabrera et al., 1992; Cabrera, Nora, and

Castaneda, 1993; Cabrera, Stampen, and Hansen, 1990; Nora and Cabrera, 1996). In the 1990s, Carebra et al. developed the model by merging two theoretical models of college persistence - Tinto's Student Integration Model (Tinto, 1975, 1982, 1988, 1987) and Bean's Student Attrition Model (Bean, 1980, 1982, 1985; Bean and Metzner, 1985). The Student Adjustment Model was initially proposed to understand the persistence of college students, and it has been used in previous studies to understand degree completion (Gross, Torres, and Zerquera, 2013). The model hypothesizes that three categories of factors may contribute to higher education outcomes, which are individual, institutional, and environmental factors. The individual factors include personal background such as gender and race/ethnicity, precollege characteristics such as academic performance in high school, academic integration such as satisfaction with course curriculum, social integration such as personal relationship with other students. The individual factors included in this study are age when matriculated, total SAT score, cumulative credits passed for GPA, count of classes with D, F, or W grade, credits registered in a Spring semester, in-state residency, gender, underrepresented minority, and first-generation college student flag. The institutional factors include institutional fit and quality such as sense of belonging. The environmental factors include financial support and encouragement from friends and family. The environmental factors included in this study are cumulative loan debt, cumulative grant aid, cumulative scholarship aid, and adjusted gross income. While it is a comprehensive framework, the Student Adjustment Model assumes the effect of financial support such as loan debt is the same within the institution, which is not practical. To better understand the effect of loan debt on graduation, this study further includes the perspective of the human capital model in the conceptual framework (Chen and Hossler, 2017).

The human capital model was developed by Gary Becker (Becker, 1964) and has been used by many studies to understand college enrollment and outcomes (Charles and Luoh, 2003; Chen and Hossler, 2017; George-Jackson, Rincon, and Martinez, 2012; Lee, 2018; Venti and Wise, 1983; Willis and Rosen, 1979). The model suggests an individual will decide to enter the labor market instead of attending school, if the marginal cost of education is higher than the expected benefit of education (Long, 2007). Loan debt increases the cost of education, and thus, potentially lowers the probability of persistence and graduation especially when the loan burden is excessive (Fossey and Bateman, 1998). On the other hand, the expected benefit of college education is different for students in different departments. According to the most recent data published by the department of education, the median earning after graduation could be less than $21,000 in some fields of study and more than $68,000 in some other fields (U.S. Department of Education, 2019). Students may feel much less financial burden from loan debt if expecting much higher future income. Therefore, the effect of loan debt on graduation can be different in different departments, so department should be included as a factor in addition to other factors included in the Student Adjustment Model.

4.4 Empirical Model

Table 4.1 shows the list of independent variables in the empirical model. The model includes eight numerical variables ($X_1, X_2, \ldots, X_8$) such as cumulative loan debt and five categorical variables ($Y_1, Y_2, \ldots, Y_5$) such as underrepresentative minority status. More details of the independent variables are described in the data and variable section.

Table 4.1: List of Independent Variables in the Model

Variable Code	Variable Description
$\mathbf{X_1}$	Cumulative loan debt
$\mathbf{X_2}$	Age during matriculation
$\mathbf{X_3}$	Total SAT score
$\mathbf{X_4}$	Cumulative credits passed for GPA
$\mathbf{X_5}$	Count of classes with D, F, or W grade
$\mathbf{X_6}$	Credits registered in spring
$\mathbf{X_7}$	Total grant aid
$\mathbf{X_8}$	Total scholarship aid
$\mathbf{Y_1}$	In-state residency
$\mathbf{Y_2}$	Gender
$\mathbf{Y_3}$	Underrepresented minority
$\mathbf{Y_4}$	First Generation College Student
$\mathbf{Y_5}$	Adjusted gross income (AGI)

The study uses a Bayesian hierarchical logit model to estimate the effects on graduation for three reasons. First, a logit model is proven to be a good method to estimate the loan effect in previous studies (Chen and Hossler, 2017; Davidson, 2015; Dowd and Coury, 2006; Dwyer, McCloud, and Hodson, 2012; Herzog, 2018; Letkiewicz et al., 2014). In a logit model, a student is assumed to graduate with a probability of θ. The logit of θ, ($log\,\frac{\theta}{1-\theta}$), is the logarithm of the odds of graduation, and the odds is the ratio of the probability of graduation over the probability of no graduation. The logit is assumed to be a linear combination of variables, as shown in the formula below. For a numerical variable X_j (j = 1, 2, …, 8), the corresponding coefficient β_j is a vector and $\beta_j[department]$ represents an element of the vector for a specific department. For example, the β for loan debt is a vector and each element in the vector represents the effect of loan debt in a department. For a categorical variable Y_k (k = 1, 2, …, 5), the corresponding coefficient α_k is a matrix and the value of each element is based on department and the value of Y_k. For example, the α for underrepresented minority is a matrix. Each element represents the effect of a specific underrepresented minority status in a specific department. Each row of the matrix represents the effects for students in a department with different underrepresented minority status. Each column represents the effects for students in different departments with the same underrepresented minority status.

$$logit(\theta) = log\frac{\theta}{1-\theta} = \sum_{j=1}^{8} \beta_j\,[department] * x_j + \sum_{k=1}^{5} \alpha_k\,[department, y_k] \quad (1)$$

Second, the uncertainty of an effect can be easily estimated by a Bayesian model (Crisp, Doran, and Reyes, 2018). The uncertainty of an effect is often characterized by the point estimate (mean or median), the standard deviation, and the corresponding credible interval. A Bayesian model assumes that an effect follows a distribution, and the characteristics of uncertainty can be directly estimated from the fitted distribution. In addition to a point estimate and a credible interval, the distribution of the effect can be visualized to show its shape and spread, so the uncertainty is well described. For example, a density plot of the fitted distribution can show the uncertainty in the effect of the loan debt.

In a Bayesian analysis, our initial uncertainty, known as a prior distribution, is modelled using the language of probability distributions. For the unknown coefficient terms of Equation (4.1), we model the initial uncertainty as using normal distributions as shown in Equation (4.2). The prior distributions are updated to posterior distributions using data observed (McElreath, 2015). The posterior distributions are the fitted distributions of the coefficients.

$$\begin{array}{r} \beta_j[department] \sim \text{Normal}(\mu_j[department], \sigma_j) \\ \alpha_k[department, y_k] \sim \text{Normal}(\mu_k[department, y_k], \sigma_k') \end{array} \tag{2}$$

Third and most importantly, a hierarchical model is used to address the small sample problem encountered because some academic departments have less than 10 students enrolled in a semester. Without Bayesian methods, estimating the effect of a factor from such a small sample size often leads to over-fitting. One way to avoid over-fitting is to assume the effect in a department is the same with the effect in the department's college where the sample size is larger. However, this strategy leads to under-fitting since the effects could be different among departments. In order to balance between over-fitting and under-fitting, a hierarchical model uses partial pooling to borrow information from an upper-level factor, when estimating the effect for a lower-level factor (McElreath, 2015). The smaller sample size the lower-level factor has, the larger influence the upper-level factor has in the estimation of the effect of the lower-level factor, and vice versa. In this case, portfolio is defined to be the upper-level for department. A portfolio is a group of departments. It is just the college of a department, except for the departments in the college of arts and sciences. The college of arts and sciences groups its departments into four portfolios. Similarly, the institution is the upper-level for portfolio. When estimating the effect of loan debt in a department, the influence of the effect in the corresponding portfolio is large if few students are enrolled in the department, and the influence is little if the department has many students.

In terms of model specification, the mean of the normal distribution ($\mu_j[department]$ in the above formula for $\beta_j[department]$) represents the mean effect of a continuous variable X_j in a department. It is drawn from $\tilde{\beta}_j$ which is a vector and represents the effect on the portfolio level. The $\tilde{\beta}_j[portfolio]$ represents the effect of a portfolio and is assumed to follow a normal distribution with a mean of $\hat{\beta}_j$. The $\hat{\beta}_j$ represents the institutional level effect and is assumed to be normally distributed. The whole hierarchical structure implements the idea that the estimation of department level effect is influenced by the portfolio level effect, and the estimation of the portfolio level effect is influenced by the institutional level effect. The assumption for the mean effect of a categorical variable Y_k in a department ($\mu_k[department, y_k]$) in the above formula for $\alpha_k[department, y_k]$ is similar to that of $\mu_j[department]$. The value of $\mu_k[department, y_k]$ is based on department and the value of Y_k. It is drawn from $\hat{\alpha}_k$ which is a vector and represents the effect only based on department. The department level effect $\hat{\alpha}_k$ is assumed to be influenced by the portfolio level effect $\hat{\beta}'_k$, and $\hat{\beta}'_k$ is assumed to follow a normal distribution. All standard deviation terms in the normal distributions are assumed to follow an exponential distribution with a parameter λ^* and λ^* assumed to follow a gamma distribution. The constant values of the parameters in the gamma distribution, exponential distribution, and normal distributions are chosen to reflect the domain knowledge that most students graduated within six years, and thus the density of the probability of graduation (θ) close to 1 should be high and the density close to 0 should be low. The model specification for the hierarchical structure is summarized below.

$$
\begin{gathered}
\lambda^* \sim \text{Gamma}(5,1) \\
\tau^* \sim \text{Exponential}({\lambda^*}^2/2) \\
where\ \lambda^*\ represents\ \lambda, \lambda', \lambda_p, \lambda_d, \lambda'_p, \lambda'_d, \\
\tau^*\ represents\ \tau, \tau', \hat{\sigma}, \sigma, \widehat{\sigma'}, \sigma', respectively
\end{gathered}
$$

$$
\begin{aligned}
\hat{\beta}_j &\sim \text{Normal}(0.4, \tau) \\
\tilde{\beta}_j[portfolio] &\sim \text{Normal}(\hat{\beta}_j, \hat{\sigma}) \\
\mu_j[department] &= \tilde{\beta}_j[portfolio] \\
\\
\hat{\beta}'_k[portfolio] &\sim \text{Normal}(0.4, \tau') \\
\hat{\beta}'_k[department] &= \hat{\beta}'_k[portfolio] \\
\hat{\alpha}_k[department] &\sim \text{Normal}(\hat{\beta}'_k[department], \hat{\sigma}') \\
\mu_k[department, y_k] &= \hat{\alpha}_k[department]
\end{aligned}
\tag{3}
$$

Figure 4.1 summarizes the Bayesian hierarchical logit model. The dependent variable, Six-year Graduation, is a binary variable following a Bernoulli distribution with a parameter θ, and θ is the probability of a student graduating within six years. The logit of θ is a linear combination of factors. The effects of the factors follow a hierarchical structure as described above.

4.5 Data and Variables

Data for this study were pulled from the enterprise data warehouse of the University of Delaware, a public research university (Carnegie classification: R1) with a population of about 18,000 undergraduate students. In fall 2009, 2010 and 2011, the university admitted 3,807, 3,366 and 3,906 (in total 11,079) first-time, full-time, first-year students, respectively. At the end of the first spring, there were 10,657 students who were still enrolled and had not graduated. Subsequently, for students who were enrolled and have not graduated, there were 9,887, 9,408, 1,859, and 369 of them at the end of the second, the third, the fourth and the fifth spring semester, respectively.

The dramatic population drop at the end of the fourth spring reflects many students had graduated in the semester. An academic year at the University of Delaware is from fall to summer. Since all data are measured at the end of students' spring semesters, as opposed to the end of an academic year which ends at the summer sessions, we will often make this explicit by including the word "Spring" in reference to each cohorts' data, e.g., data collected for first-year students are referred to the first spring.

The dependent variable in a model is whether a student graduated within six years since matriculation. For students who were enrolled and had not graduated at the end of the first spring, 9,098 or 85.4% of them eventually graduated within six years. For students who were enrolled and had not graduated at the end the following four springs, 91.5%, 95.2%, 83%, and 58% eventually graduated within six years, as shown in Table 4.2.

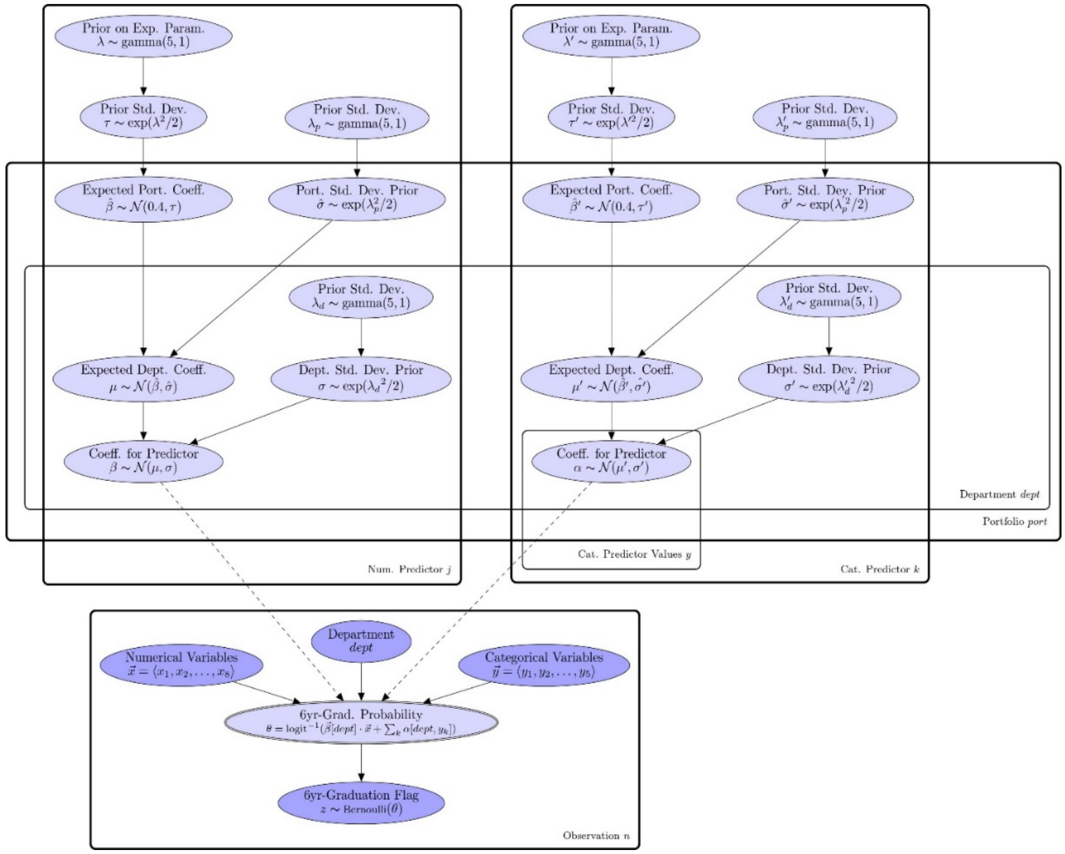

Figure 4.1: The Illustration of the Bayesian Hierarchical Logit Model

Table 4.2: Summary of the Enrollment Headcounts, Six-Year Graduation Headcounts and Six-Year Graduation Rates for Students Who Were Still Enrolled and Had Not Graduated at the End of Each Spring

	Enrollment headcount	Graduated within 6 years	% Graduated within 6 years
End of Spring 1	10,657	9,098	85.4
End of Spring 2	9,887	9,044	91.5
End of Spring 3	9,408	8,955	95.2
End of Spring 4	1,859	1,543	83.0
End of Spring 5	369	214	58.0

Loan debt is the focal independent variable in this study. The cumulative loan debt of a student is computed at the end of each spring semester. Table 4.3 shows the summarized information of cumulative loan debt. Students who did not borrow loans were excluded in the table. For students who were enrolled at the end of the first spring, 5,875 or 55.1% of students accumulated loan debt. The overall average loan debt is $7,643, and it is $7,756 and $7,620, respectively, for those who do not graduate within six years and those who eventually graduate within six years. The headcount of students with loan debt decreases slightly from the first spring to the third spring, and then dramatically decreases at the end of the fourth spring due to graduation. The percentage of students with loan debt increases slowly each year. The overall average of cumulative loan debt increases each year for more than $5,000 in the first four years. The average loan debt of students who do not graduate within six years is higher than the overall average in the 1st year and the 4th year, and it is higher for students who graduate within six years in other years.

Table 4.3: Headcount, Percentage and Average Loan Debt of Students Who Were Enrolled, Had Not Graduated and Accumulated Loan Debts at the End of Each Spring Semester

	Headcount	Percentage	Mean (All Enrolled)	Mean (Not Graduated 6 within years)	Mean (Graduated within 6 years)
End of Spring 1	5,875	55.1	$7,643	$7,756	$7,620
End of Spring 2	5,697	57.6	$15,849	$15,811	$15,854
End of Spring 3	5,609	59.6	$24,502	$22,953	$24,588
End of Spring 4	1,165	62.7	$30,632	$31,426	$30,460
End of Spring 5	258	69.9	$32,433	$32,050	$32,731

The loan debt effect is investigated by department. A student's department is where the student was enrolled in a semester. There are 53 departments in total. Table 4.4 shows the headcounts of students who were enrolled and had not graduated at the end of each spring by department. The department names are replaced with unique numbers for privacy consideration, and similarly the headcounts are masked if enrollment is under 10. Some departments were more popular than the others. For instance, hundreds of students were enrolled in the department 0, while few students were enrolled in the department 1.

Table 4.4: Headcounts of Students Who Were Enrolled and Had Not Graduated at the End of Each Spring by Department

	End of Spring 1	End of Spring 2	End of Spring 3	End of Spring 4	End of Spring 5
0	445	567	606	104	22
1	Privacy Suppressed	Privacy Suppressed	Privacy Suppressed	Privacy Suppressed	Privacy Suppressed
2	11	19	30	Privacy Suppressed	Privacy Suppressed
3	223	187	176	29	Privacy Suppressed
4	45	35	34	6	Privacy Suppressed
5	48	63	72	18	Privacy Suppressed
6	132	131	118	34	Privacy Suppressed
7	21	20	20	Privacy Suppressed	Privacy Suppressed
8	18	22	22	Privacy Suppressed	Privacy Suppressed
9	266	368	374	62	10
10	281	528	541	107	16
11	260	324	335	99	22
12	55	97	119	35	Privacy Suppressed
13	612	436	377	65	13
14	92	97	89	18	Privacy Suppressed

15	19	15	12	Privacy Suppressed	Privacy Suppressed
16	312	248	215	40	Privacy Suppressed
17	156	137	127	18	Privacy Suppressed
18	392	375	340	66	11
19	351	387	324	31	Privacy Suppressed
20	127	123	114	42	Privacy Suppressed
21	77	164	186	68	12
22	171	142	131	35	10
23	18	Privacy Suppressed	Privacy Suppressed	Privacy Suppressed	Privacy Suppressed
24	241	275	288	67	18
25	72	82	78	17	Privacy Suppressed
26	24	32	30	Privacy Suppressed	Privacy Suppressed
27	221	218	207	20	Privacy Suppressed
28	149	148	135	29	Privacy Suppressed
29	24	26	26	10	Privacy Suppressed
30	26	32	30	Privacy Suppressed	Privacy Suppressed
31	187	198	201	65	16
32	251	323	323	63	Privacy Suppressed
33	223	277	296	56	16
34	479	499	486	59	Privacy Suppressed
35	123	127	133	53	12
36	305	19	Privacy Suppressed	Privacy Suppressed	Privacy Suppressed
37	98	133	135	Privacy Suppressed	Privacy Suppressed
38	181	150	137	36	Privacy

					Suppressed
39	390	325	297	56	12
40	75	80	82	18	Privacy Suppressed
41	140	133	126	18	Privacy Suppressed
42	22	18	15	Privacy Suppressed	Privacy Suppressed
43	39	38	32	12	Privacy Suppressed
44	24	23	25	11	Privacy Suppressed
45	373	369	354	58	15
46	547	515	463	91	13
47	376	384	375	45	Privacy Suppressed
48	21	18	19	Privacy Suppressed	Privacy Suppressed
49	380	408	389	48	Privacy Suppressed
50	224	280	290	81	14
51	1305	258	51	14	Privacy Suppressed
52	Privacy Suppressed	Privacy Suppressed	11	Privacy Suppressed	Privacy Suppressed

The selection of control variables is based on prior research, domain knowledge of subject experts and data availability. Total SAT score serves as a proxy of pre-college academic preparation. The total SAT score is the highest total SAT score if a student took SAT several times, and it could be converted from composite ACT score if a student submitted ACT scores. There are 324 students without a total SAT score and are excluded from the study. Most of them are international students.

In-state residency, gender, first generation college, and underrepresented minority serve as demographic background. They are dichotomous variables. In-state residency indicates whether a student was a Delawarean resident (coded 1) and thus paid in-state tuition rate. Female is coded 0 and male is coded 1 for gender. First generation college indicates whether a student is a first-generation college student (coded 1) according to the highest education level of the student's parents. The definition of underrepresented minority can vary among institutions, and this study uses the definition at the University of Delaware. Underrepresented minority flag is coded 1 if a student had an ethnicity of Black, Hispanic, Native American Indian, Hawaiian, or other Pacific Islander.

Adjusted gross income (AGI), cumulative grant aid, and cumulative scholarship aid serve as students' financial background. The AGI is the sum of the AGI of a student and the parents if the student is a dependent, and it is the AGI of the student if he/she is independent. AGI is grouped into four categories, AGI unknown (AGI is missing), Low AGI (AGI < 61K), Medium AGI (AGI is less than the 3rd quartile) and high AGI (AGI is larger than or equal to the 3rd quartile). Grant aid is typically need-based aid, and scholarship aid is typically merit-based aid. Like the cumulative loan debt, the cumulative grant aid and cumulative scholarship aid are computed as of the end of each spring.

Count of classes with DFW grade (D or F grade, or withdraw), cumulative credits passed for GPA and credits registered in the current semester serve as college academic measures. A student needs to retake a class with D, F or W grade, if a class is a pre-requisite of another class or due to the departmental requirement, and thus the progress towards graduation is potentially delayed. Cumulative credits passed for GPA represents the overall progress towards graduation. Credits registered is the number of credits attempted in a spring, indicating a student's effort/attitude toward graduation. It should be noted that cumulative GPA is not included, as in the study by Jones-White et al. (2014). Although GPA is a strong factor on academic performance, it does not directly represent the progress toward graduation. The DFW count and cumulative credits better capture the progress.

The summary statistics of the control variables are shown in Table 4.5, for students who were enrolled and had not graduated at the end of the first spring. The average SAT score was 1288. About one third of the students were in-state (32%). There were fewer male students (42%) than female. There are 13% and 11% of students who are first generation college students and underrepresented minority students, respectively. On average, a student passed 29 credits by the end of the spring, with only 0.33 classes with a DFW grade. One student accumulated 112 credits in the first year, because the student was a long-time continuing education student in the university and transferred all the credits after being admitted as a full-time, first-time, first-year degree-seeking student. Students tend to register 15 credits in the spring. We did not know 21% students' AGI, 14% of the students had low AGI (less than $61,000), 47% of them had medium AGI (between $61,000 and third quartile), and the rest 18% students had high AGI (above third quartile). On average the students were awarded $1.93K and $2.85K from grant aid and scholarship aid, respectively, including students who did not obtain any grant or scholarship aid. The summary statistics for other springs are in Appendix D.

Table 4.5: Descriptive Statistics of Student Characteristics at the End of the First Spring

	Mean	SD	Min	Max
Total SAT score	1,288	122	740	1,590

Age when matriculation	18	0.60	16	35
In-state Residency	0.32		0	1
Male	0.42		0	1
First Generation College Student	0.13		0	1
Underrepresented minority Student	0.11		0	1
Cumulative credits passed for GPA	29	5.16	0	112
Count of classes with DFW grade	0.33	0.77	0	7
Credits registered in a spring	15	1.64	0	22
AGI unknown	0.21		0	1
Low AGI	0.14		0	1
Medium AGI	0.47		0	1
High AGI	0.18		0	1
Total grant aid (in $1,000s)	1.93	3.97	0	24.24
Total scholarship aid (in $1,000s)	2.85	5.59	0	45.8

All numeric factors, namely cumulative loan debt, total SAT score, cumulative credits passed for GPA, count of classes with DFW grade, credits registered in a spring, cumulative grant aid and cumulative scholarship aid, are standardized before using as input of a model. Specifically, they are subtracted by mean and then divided by standard deviation. Standardization brings the factors to the same scale, in order to avoid biased estimation of the coefficients in a model (McElreath, 2015).

4.6 Results and Discussion

The posterior distributions of the effects were estimated in RStudio (version 1.2.1335) using the greta package (version 0.2.5) (Golding, 2019) and TensorFlow Probability (0.5.0) (Abadi et al., 2016). The effects of variables are estimated using four Markov chains. Each chain contains a warm-up period of 3,000 steps followed by a sampling period of 1,000 samples. In total, 4,000 samples are drawn from posterior distributions.

The results are interpreted in terms of odds ratio. In this study, odds are the probability of six-year graduation over the probability of not graduating within six years, and the odds ratio is the ratio of two different odds. An odds ratio larger than 1 indicates a positive effect on the six-year graduation, while an odds ratio less than 1 indicates a detrimental effect. Given the research questions focus on the loan debt effect on six-year graduation, the discussion is mainly limited to the results of loan debt effect. The comprehensive results including odds ratios of all factors by department are available upon request.

4.6.1 The First Spring

Loan debt is a detrimental effect on six-year graduation on the institutional level for students who were active and had not graduated at the end of the first spring. Figure 4.2 shows the 90% credible intervals of the odds ratios for the factor effects on the institutional level, where department is not considered in the model. The effects of loan debt, grant aid, and merit aid are calculated with respect to each additional $1,000. The vertical line represents an odds ratio of 1. The odds ratio for loan debt is statistically less than 1 (odds ratio: median = 0.98, 5% quantile = 0.97, 95% quantile = 0.99). If we further assume the probability of six-year graduation without loan debt is 80%, the probability decreases to 79.7%, with the additional $1,000 loan debt and holding other factors constant. Based on the institutional level results, the university could design a universal financial aid policy to help students with loan debt, in order to reduce their financial burden and increase their change to graduate within six years. However, the following results show the financial aid policy can be more efficient if considering the department level effect.

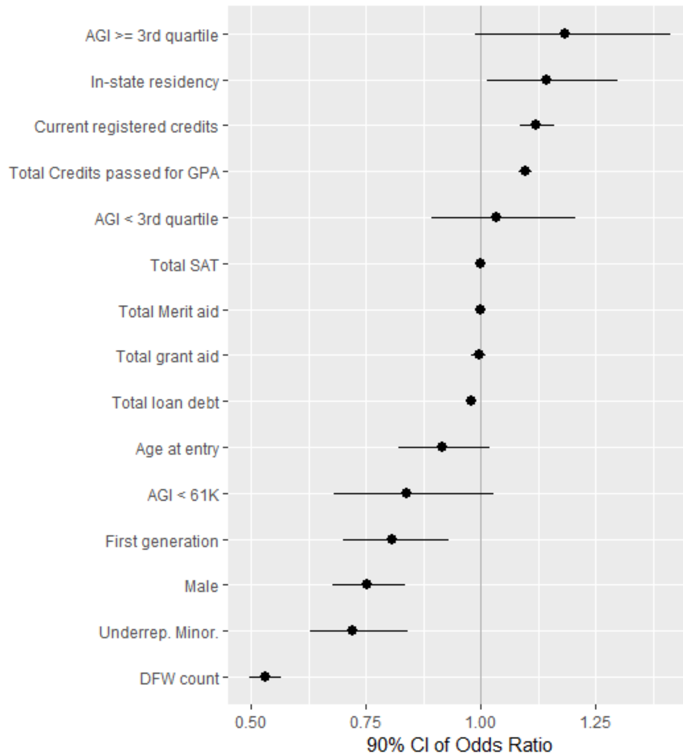

Figure 4.2: The 90% Credible Intervals of the Odds Ratios for the Factor Effects on Institutional Level for the First Spring

Loan debt has distinctively different effects among departments for students who were active and had not graduated at the end of the first spring. The difference is described using the following three figures, Figure 4.3, Figure 4.4 and Figure 4.5. Figure 4.3 shows the 90% credible intervals of the odds ratios for the loan debt effect by department, with respect to each additional $1,000 loan debt. The vertical line represents an odds ratio of 1. The odds ratio is statistically less than 1 for fifteen of the fifty-two departments, for example, department 35 (odds ratio: median = 0.93, 5% quantile = 0.9, 95% quantile = 0.96). Specifically, the odds ratio of six-year graduation tends to be 0.93 for students in the department 35 with each additional $1,000 loan debt, holding other factors constant. If we further assume the probability of six-year graduation without loan debt is 80%, the probability decreases to 78.8%, with the additional $1,000 loan debt. Note that the decrease of the probability is more severe for the department, compared to the overall institution. On the other hand, the credible intervals of the odds ratios span across 1 for thirty-seven departments, for example, department 47, indicating the loan debt is neither a positive nor a detrimental effect for the students in the department.

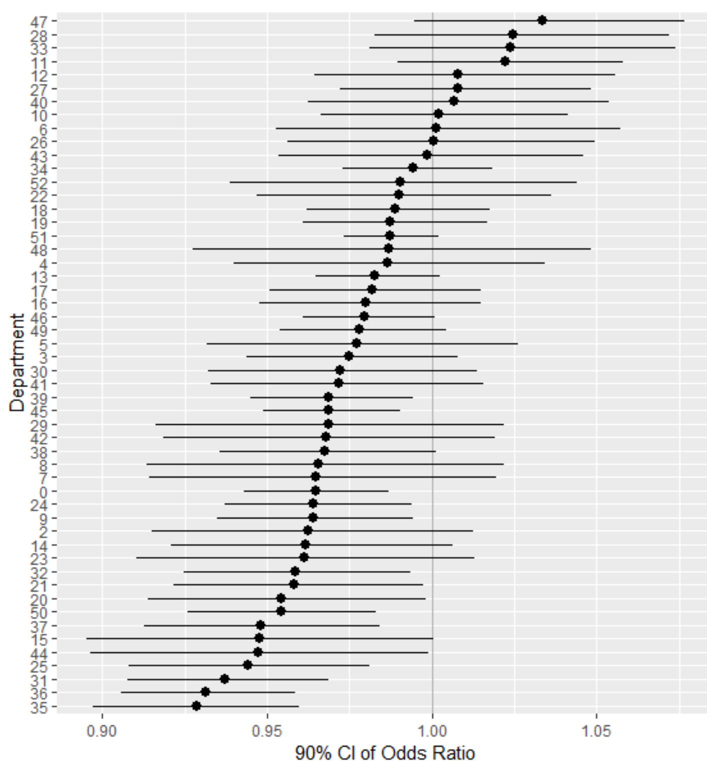

Figure 4.3 The 90% Credible Intervals of the Odds Ratios for the Loan Debt Effect by Department for the First Spring

The difference of the loan debt effects between departments can also be viewed by the density plots of the posterior distributions of the odds ratios. For example, Figure 4.4 shows the density plots of the odds ratios of department 35 and department 47. The two density plots barely overlap, indicating the loan debt has distinctively different effects in the two departments. Moreover, the density plots agree with the 90% credible intervals of the odds ratios. The estimated odds ratios for department 35 are always smaller than 1, indicating a detrimental effect for six-year graduation. On the other hand, although the estimated odds ratios are more likely to be larger than 1 for department 47, the proportion of estimated odds ratios less than 1 is not neglectable, and thus we cannot conclude the loan debt is a positive or detrimental effect on six-year graduation.

If choosing the loan debt effect in department 35 as reference, students in twenty-six departments were affected differently. Figure 4.5 shows the 90% credible intervals of the odds ratios of the loan debt effect, compared with the department 35. Here an odds ratio larger than 1 indicates the loan debt effect is more positive or less detrimental in a department than that in the department 35. For example, the odds ratio tends to be 1.11 between department 47 and department 35, indicating students in the department 47 are less detrimentally affected by loan debt than those in the department 35. On the other hand, the credible intervals span across 1 for twenty-five departments, indicating the loan debt has similar effects in the departments with that in the department 35.

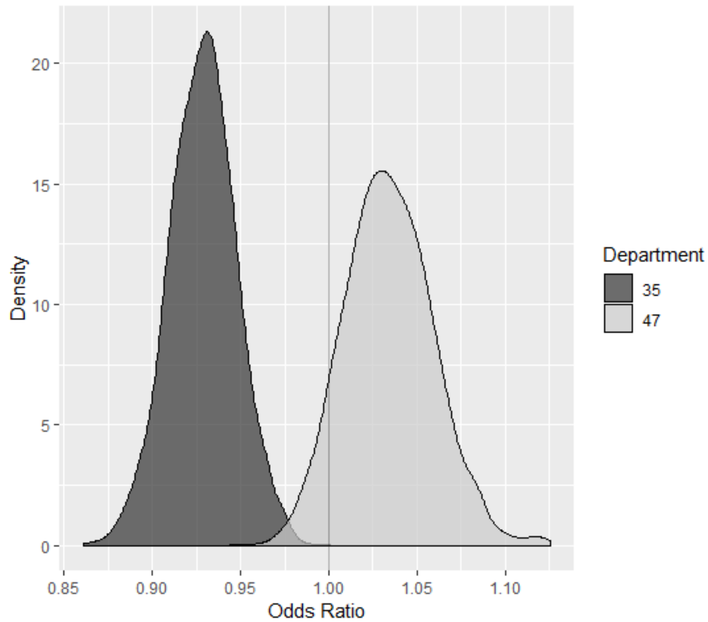

Figure 4.4: The Density Plots of the Odds Ratios of Loan Debt Effect in Department 35 And Department 47 for the First Spring

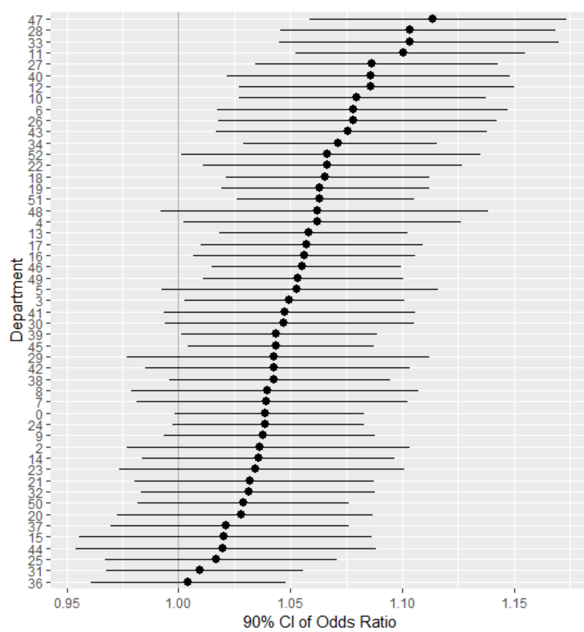

Figure 4.5: The 90% Credible Intervals of the Odds Ratios of Loan Debt Effect for the First Spring, Using Department 35 as Reference

Based on the results above, the university should design a differential financial aid policy by department for the first-year students, along with other existing factors when awarding financial aid packages. The university could offer more institutional scholarship aid to the students in the fifteen departments where the loan debt is a significant detrimental effect on six-year graduation, to slow the accumulation of loan debt. From the students' perspective, the differential financial aid policy more efficiently reduces their financial burden, and thus helps them focus more on academic progress and the continuation of the degree. From the institutional perspective, the policy helps to enhance the overall six-year graduation rate, which is a critical performance measurement of a four-year institution.

4.6.2 The Second Spring

Similar to the effect for the first-year students, loan debt is still a detrimental effect on six-year graduation on the institutional level for students who remained enrolled and had not graduated at the end of the second spring. Figure 4.6 shows the 90% credible intervals of the odds ratios for the factor effects on the institutional level. The odds ratio for loan debt is statistically less than 1 (odds ratio: median = 0.97, 5% quantile = 0.98, 95% quantile = 1.00). If we further assume the probability of six-year graduation without loan debt is 80%, the probability decreases to 79.8%, with the additional $1,000 loan debt and holding other factors constant. Without considering the department effect, it is reasonable to conclude more financial aid in terms of grant and scholarship should be awarded to students with loan debt, in order to reduce the detrimental effect of loan debt. However, the following results show this is not an efficient policy.

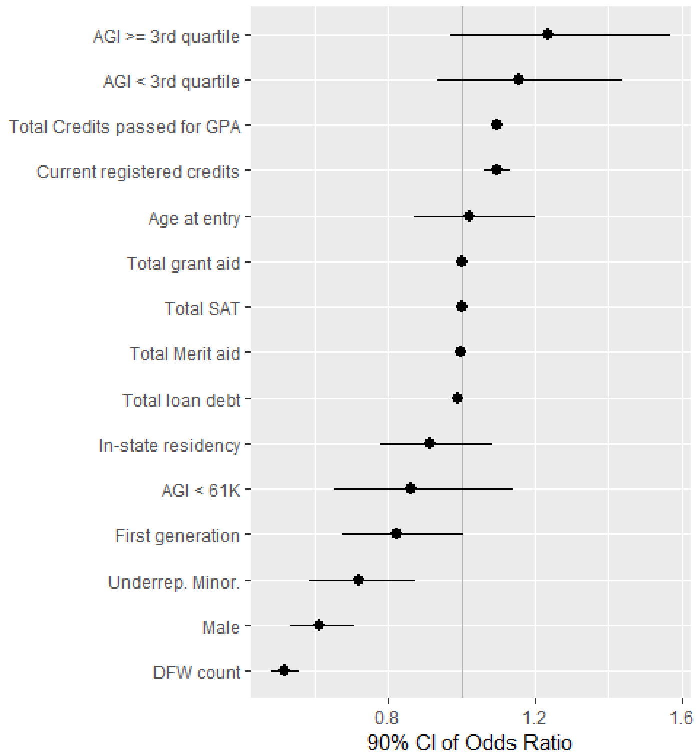

Figure 4.6: The 90% Credible Intervals of the Odds Ratios for the Factor Effects on Institutional Level for the Second Spring

Loan debt is generally not a detrimental effect for the students who remained enrolled and had not graduated at the end of the second spring, according to Figure 4.7. Figure 4.7 shows the 90% credible intervals of the odds ratios for the loan debt effect by department, with respect to each additional $1,000 loan debt. The odds ratio is statistically less than 1 for only three departments, for example, department 36 (odds ratio: median = 0.97, 5% quantile = 0.95, 95% quantile = 0.99). The credible intervals of other fifty departments span across 1, indicating the loan debt does not directly affect the six-year graduation.

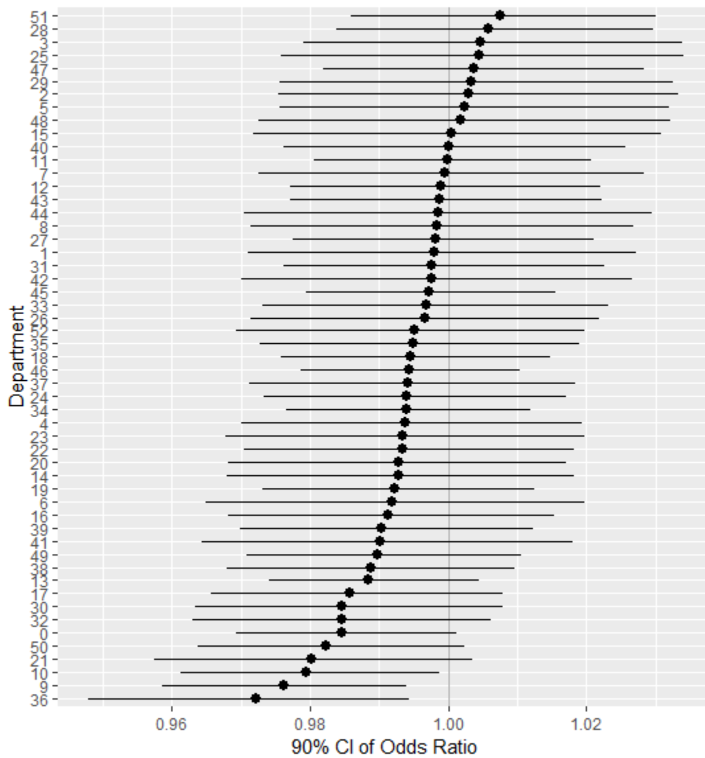

Figure 4.7: The 90% Credible Intervals of the Odds Ratios of Loan Debt Effect by Department for the Second Spring

Moreover, loan debt does not have distinctively different effects among departments for second-year students. The 90% credible intervals of the odds ratios overlap with each other, indicating similar loan debt effects among departments. The difference of the loan debt effect between department 35 and department 47 is shown again using density plots in Figure 4.8. Different from the first spring, the two density plots of odds ratios almost overlay with each other, indicating the loan debt has very similar effects in the two departments. However, we caution that the results may change if the institution applies differential financial aid policy by department for the first-year students. For example, the extra financial aid to the first-year students in department 35 helps some students remain enrolled, who would stop pursuing a degree otherwise, but they would face the delayed financial burden if the extra help is removed in their second year. As a result, we propose the university to use a differential financial aid policy for the first-year students, and then use randomized experimental design to investigate whether similar differential policy should still be applied for the second-year students.

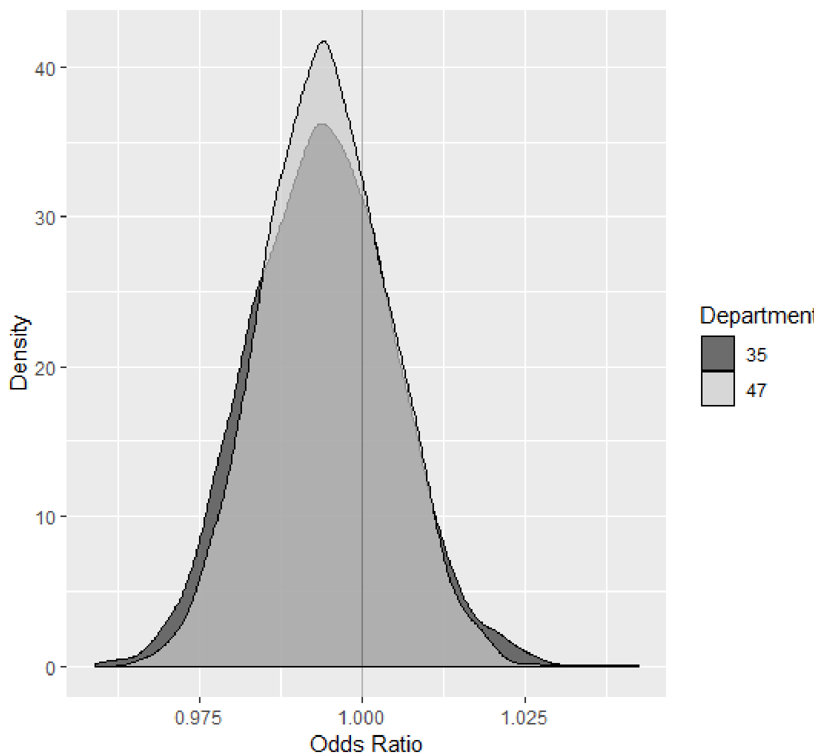

Figure 4.8: The Density Plots of the Odds Ratios of Loan Debt Effect in Department 35 and Department 47 for the Second Spring

4.6.3 The Third, the Fourth, and the Fifth Spring

Different from the first and the second springs, loan debt does not show detrimental effect on six-year graduation anymore for students who remained enrolled but had not graduated at the end of the third, the fourth, and the fifth spring. The graphs of the 90% credible interval of odds ratio of the factors for the three springs are shown in Appendix E. The credible intervals for the loan debt effect span across 1, indicating the loan debt does not directly affect six-year graduation.

The departmental level effects agree with the institutional level effects for the same group of students. The graphs of 90% credible intervals of the odds ratios of loan debt effect by department are in Appendix E. All departments' 90% credible intervals for the loan debt effect span across 1, indicating the loan debt does not directly affect six-year graduation. Loan debt shows similar effects on six-year graduation among departments since the intervals overlap with each other. Like the argument for the second year's results, if the university starts to implement differential financial aid policy by department for the first-year students, loan debt can still be a detrimental effect in later years if extra aid is not awarded in the later years. Again, the university needs to investigate whether extra aid to some departments' students will just delay the detrimental effect of loan debt.

4.6.4 Change of Loan Debt Effect Over Years

For thirteen departments, loan debt changes from detrimental effect in the first spring to indirect effect in the second spring on six-year graduation. Figure 4.9 shows the change of the 90% credible intervals of the odds ratios from six of the departments over years. The odds ratios for the first spring are statistically less than 1 according to the intervals, indicating a detrimental effect. On the other hand, the intervals span across 1 for the other springs, indicating an indirect effect.

For other department, loan debt does not directly affect six-year graduation for all years, with three exceptions (results available upon request). The loan debt remains as a detrimental effect in the second spring for department 9 and department 36, and then becomes a non-significant effect afterwards. The loan debt changes from a non-significant effect in the first spring to a detrimental effect in the second spring for department 10, and then changes back to a non-significant effect afterwards.

Therefore, the effect of loan debt on graduation changes over years in the same department. In many cases it changes from a detrimental effect in the first spring to a non-significant effect afterwards. Similar to the earlier proposition, the university should focus on helping the first-year students to slow their debt accumulation, since a differential policy by department should utilize the financial resource efficiently. The university needs to investigate whether the differential is needed in the following years.

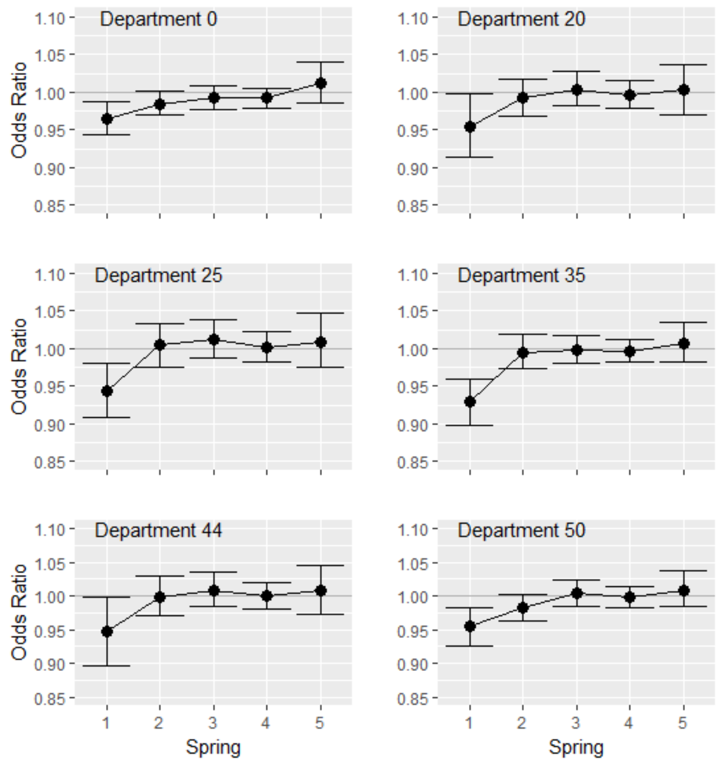

Figure 4.9: The Change of the 90% Credible Intervals of the Odds Ratios of Loan Debt Effect Over Years in Selected Departments

4.7 Conclusion

Bayesian hierarchical models are useful tools to estimate the effects of loan debt on the six-year graduation by department. The hierarchical structure uses partial pooling technique to estimate the department level effects. More specifically, the estimated portfolio level effects influence the estimation of the department level effects. The less students enrolled in a department, the more the estimated department level effect is influenced by the estimated portfolio level effect, and vice versa. This strategy was particularly critical for departments with low enrollment to avoid the over-fitting issue due to small sample size.

Loan debt affects six-year graduation differently for the first-year students in different departments. The loan debt is a detrimental factor in some departments, while it has no direct effect in the others. The university is currently using the universal financial aid policy for all departments. The results suggest that a differential policy with department as one of the factors will be more useful to help students graduate on time. The accumulation of loan debt can be slowed by offering more institutional scholarships to the first-year students. For institutions with increasing institutional financial aid budget over years, the proposal should lead to a faster increase of institutional financial aid for departments where students tend to suffer more from loan debt than the other departments, while it should not lead to a decrease in the other departments. For institutions with stable institutional financial aid budgets and abundant overall institutional budgets, the proposed policy suggests to increase the institutional financial aid budget, so the strategy proposed for the first case can be used. On the other hand, caution should be taken when the policy leads to a redistribution of a finite institutional financial aid budget, because the enrollment could hurt departments that do not benefit from this policy. Moreover, an institution should pay attention to potential equity impacts when adjusting the institutional financial aid policy, particularly for underserved departments or students.

Conversely from the first-year students, the study finds that loan debt has similar effects on six-year graduation rates among departments for the second-year, the third-year, the fourth-year, and the fifth-year students. The loan debt effect typically becomes or remains non-significant on six-year graduation for the second-year students and thereafter. Therefore, students who remain enrolled in the following years become insensitive to loan debt as long as they overcome the financial stress from loan debt in the first year, under the current universal financial aid policy. The findings help to explain the inconclusive relationship between loan debt and on time graduation in previous research, since the relationship can be different between different student populations, even for students in the same university but at different academic stages. If the differential policy is implemented, students in some departments will receive extra scholarship to slow the accumulation of loan debt, and thus less financial burden helps them to remain enrolled and continue to pursue a degree. The university needs to investigate whether the differential policy is needed for the following years in case the extra aid in the first year just delays the detrimental effect of loan debt. Moreover, the policy may encourage increased enrollments in departments who benefit from it, but institutions should monitor whether the proposed policy affects the pattern of the major change, if the policy is only applied to the first-year students. We hope this study will be a step to future experiments and research on differential financial aid policy.

Chapter 5

CONCLUDING REMARKS AND FUTURE DIRECTIONS

In this dissertation, we have delved into the realm of customer acquisition and retention within the higher education sector, primarily focusing on innovative approaches to student enrollment management. Our research has aimed to provide valuable insights and recommendations for the strategic refinement of student admission, retention, and graduation processes.

In Chapter 2, we embarked on an exploration of students' deposit decisions as a dynamic and evolving process. Employing piecewise exponential models enriched with Bayesian hierarchical structures, our goal was to strike a balance between underfitting and overfitting while capturing the time-varying effects of factors influencing deposit decisions. Our findings shed light on whether these factors exhibit consistency throughout the year or undergo changes from one academic year to the next. Moreover, we presented actionable insights to the Admission Office at UD, allowing them to gain a deeper understanding of distinct behaviors exhibited by different student demographics, such as gender and Pell grant eligibility. We also examined how financial aid offerings and campus events impact deposit decisions.

In Chapter 3, we delved into the modeling of enrolled students' academic journeys, acknowledging the potential for dropout stemming from academic, economic, or social factors. Utilizing structural neural networks within piecewise exponential models, we not only enhanced our predictive capabilities for dropout risks but also gained valuable insights into the relative importance of these factors. Our findings revealed that academic reasons predominantly influence dropout rates among students.

Chapter 4 focused on investigating the impact of student loan debt on the timely graduation of students across various academic departments. Given the variability in departmental sizes, we introduced a Bayesian hierarchical structure into our logit models to mitigate potential overfitting. Our analysis demonstrated that the effect of loan debt varies by department, with significant implications, especially for first-year students.

As we conclude this dissertation, we recognize that the field of student enrollment management offers numerous avenues for further exploration. Beyond the scope of this work, we outline several promising directions:

Early-Stage Analysis: Consider initiating the analysis as early as the prospect and inquiry stages to inform recruitment and marketing strategies, thereby shaping a more desirable applicant pool.

Retention Strategies: Explore the modeling of different types of dropout scenarios, distinguishing between students who leave higher education entirely and those who transfer to other institutions. This differentiation can facilitate the development of tailored retention strategies.

Alumni Engagement: Extend the analysis beyond graduation to encompass alumni engagement. By enhancing alumni giving rates, institutions can bolster their financial health.

In closing, this dissertation serves as a stepping stone into the ever-evolving landscape of student enrollment management. We hope that our research contributes to the ongoing enhancement of customer acquisition and retention in the higher education sector and inspires further investigations into this dynamic field.

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–283.

Adams-Johnson, S., Cranmore, J., Holloway, A. M. & Wiley, J. D. (2019). Higher education recruitment in the united states: A chronology of significant literature. *Journal of Educational Administration and History*, *51*(3), 213–238.

Albreiki, B., Zaki, N. & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552.

Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. *International Journal of Modern Education and Computer Science*, *9*(8), 9.

Ameri, S., Fard, M. J., Chinnam, R. B. & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 903–912.

Aulck, L., Nambi, D., Velagapudi, N., Blumenstock, J. & West, J. (2019). Mining university registrar records to predict first-year undergraduate attrition. *In Proceedings of the 12th International Conference on Educational Data Mining*, 9–18.

Aulck, L., Velagapudi, N., Blumenstock, J. & West, J. (2016). Predicting student dropout in higher education. *arXiv Preprint arXiv:1606.06364*.

Austin, P. C. (2017). A tutorial on multilevel survival analysis: Methods, models and applications. *International Statistical Review*, *85*(2), 185–203.

Baker, R. S., Yacef, K., et al. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3–17.

Bean, J. P. (1980). Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, *12*(2), 155–187.

Bean, J. P. (1982). Student attrition, intentions, and confidence: Interaction effects in a path model. *Research in Higher Education*, *17*(4), 291–320.

Bean, J. P. (1985). Interaction effects based on class level in an explanatory model of college student dropout syndrome. *American Educational Research Journal*, *22*(1), 35–64.

Bean, J. P. & Metzner, B. S. (1985). A conceptual model of nontraditional undergraduate student attrition. *Review of Educational Research*, *55*(4), 485–540.

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago press.

Becker, G. S. (2009). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago press.

Berens, J., Schneider, K., Görtz, S., Oster, S. & Burghoff, J. (2019). Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods. *Journal of Educational Data Mining*, *11*(3), 1–41.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv Preprint arXiv:1701.02434*.

Bettinger, E. P., Long, B. T., Oreopoulos, P. & Sanbonmatsu, L. (2012). The role of application assistance and information in college decisions: Results from the h&r block FAFSA experiment. *The Quarterly Journal of Economics*, *127*(3), 1205–1242.

Bouchrika, I. (2023). *College dropout rates: 2023 statistics by race, gender & income*. https://research.com/universities-colleges/college-dropout-rates

Bowen, W. G., Chingos, M. M. & McPherson, M. S. (2009). *Crossing the finish line: Completing college at america's public universities* (Vol. 52). Princeton University Press.

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

Burns, R., & Webber, K. L. (2019). Achieving the promise of educational opportunity: Graduate student debt for STEM vs. non-STEM students, 2012. *Journal of Student Financial Aid*, *48*(3), 5.

Cabrera, A. F., Castaneda, M. B., Nora, A. & Hengstler, D. (1992). The convergence between two theories of college persistence. *The Journal of Higher Education*, *63*(2), 143–164.

Cabrera, A. F., Nora, A. & Castaneda, M. B. (1993). College persistence: Structural equations modeling test of an integrated model of student retention. *The Journal of Higher Education*, *64*(2), 123–139.

Cabrera, A. F., Stampen, J. O. & Hansen, W. L. (1990). Exploring the effects of ability to pay on persistence in college. *The Review of Higher Education*, *13*(3), 303–336.

Cai, C. & Fleischhacker, A. (2022). The effect of loan debt on graduation by department: A bayesian hierarchical approach. *Journal of Student Financial Aid*, *51*(2), 5.

Cannistrà, M., Masci, C., Ieva, F., Agasisti, T. & Paganoni, A. M. (2022). Early-predicting dropout of university students: An application of innovative multilevel machine learning and statistical techniques. *Studies in Higher Education*, *47*(9), 1935–1956.

Carpenter, P. G. & Fleishman, J. A. (1987). Linking intentions and behavior: Australian students' college plans and college attendance. *American Educational Research Journal*, *24*(1), 79–105.

Carroll, D. (1987). Student financial assistance and consequences. *147th Annual Meeting of the American Statistical Association, San Francisco*.

Chapman, R. G. (1979). Pricing policy and the college choice process. *Research in Higher Education*, *10*(1), 37–57.

Charles, K. K. & Luoh, M.-C. (2003). Gender differences in completed schooling. *Review of Economics and Statistics*, *85*(3), 559–577.

Chen, J. & Hossler, D. (2017). The effects of financial aid on college success of two-year beginning nontraditional students. *Research in Higher Education*, *58*(1), 40–76.

College Board. (2020). *Average rates of growth of published charges by decade*. https://trends.collegeboard.org/college-pricing/figures-tables/average-rates-growth-published-charges-decade.

Conklin, M. E. & Dailey, A. R. (1981). Does consistency of parental educational encouragement matter for secondary school students? *Sociology of Education*, 254–262.

Crisp, G., Doran, E. & Reyes, N. A. S. (2018). Predicting graduation rates at 4-year broad access institutions using a bayesian modeling approach. *Research in Higher Education*, *59*(2), 133–155.

Davidson, J. C. (2015). Examining zero expected family contribution as a new criterion for “low income”: comparing the impact on student persistence at two-and four-year institutions. *Community College Journal of Research and Practice*, *39*(5), 442-460.

Davidson, J. C., & Holbrook, W. T. (2014). Predicting persistence for first-time undergraduate adult students at four-year institutions using first-term academic behaviors and outcomes. *The Journal of Continuing Higher Education*, *62*(2), 78-89.

Davies, M. & Kandel, D. B. (1981). Parental and peer influences on adolescents' educational plans: Some further evidence. *American Journal of Sociology*, *87*(2), 363–387.

DesJardins, S. L. et al. (1994). *Studying the determinants of student stopout: Identifying "true" from spurious time-varying effects.*

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

Dowd, A. C. & Coury, T. (2006). The effect of loans on the persistence and attainment of community college students. *Research in Higher Education*, *47*(1), 33–62.

Durkheim, E. (1897). *Le suicide: étude de sociologie*. Alcan.

Dwyer, R. E., McCloud, L. & Hodson, R. (2012). Debt and graduation from American universities. *Social Forces*, *90*(4), 1133–1155.

Dynarski, S. & Scott-Clayton, J. (2013). *Financial aid policy: Lessons from research*. National Bureau of Economic Research.

Fan, J., Ke, Z. T., Liao, Y. & Neuhierl, A. (2022). Structural deep learning in conditional asset pricing. https://ssrn.com/abstract=4117882 or http://dx.doi.org/10.2139/ssrn.4117882.

Fossey, R. & Bateman, M. (1998). *Condemning students to debt: College loans and public policy.* ERIC.

Franke, R. (2019). Take it, or leave it? Analyzing how unsubsidized federal loans affect six-year degree attainment across income groups. *Journal of Student Financial Aid*, *48*(3), 2.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, *10*(1), 101–113.

George-Jackson, C. E., Rincon, B., & Martinez, M. G. (2012). Low-income engineering students: Considering financial aid and differential tuition. *Journal of Student Financial Aid*, *42*(2), 1.

Gilmour Jr, J. E. et al. (1981). *How high school students select a college.*

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

Golding, N. (2019). Greta: Simple and scalable statistical modelling in r. *Journal of Open Source Software*, *4*(40), 1601.

Gross, J. P., Cekic, O., Hossler, D., & Hillman, N. (2009). What Matters in Student Loan Default: A Review of the Research Literature. *Journal of Student Financial Aid*, *39*(1), 19-29.

Gross, J. P., Hossler, D., & Ziskin, M. (2007). Institutional aid and student persistence: An analysis of the effects of institutional financial aid at public four-year institutions. *Journal of Student Financial Aid*, *37*(1), 6.

Gross, J. P., Torres, V. & Zerquera, D. (2013). Financial aid and attainment among students in a state with changing demographics. *Research in Higher Education*, *54*(4), 383–406.

Harrast, S. A. (2004). Undergraduate Borrowing: A Study of Debtor Students and Their Ability to Retire Undergraduate Loans. *Journal of Student Financial Aid*, *34*(1), 21-37.

Heller, D. E. (2008). The impact of student loans on college access. *The Effectiveness of Student Aid Policies: What the Research Tells Us*, 39–68.

Hemelt, S. W., Stange, K. M., Furquim, F., Simon, A. & Sawyer, J. E. (2018). *Why is math cheaper than english? Understanding cost differences in higher education* (Working Paper No. 25314; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w25314

Heredia-Jimenez, V., Jimenez, A., Ortiz-Rojas, M., Marín, J. I., Moreno-Marcos, P. M., Munoz-Merino, P. J. & Kloos, C. D. (2020). An early warning dropout model in higher education degree programs: A case study in ecuador. *LAUR@ EC-TEL*, 58–67.

Herzog, S. (2018). Financial aid and college persistence: Do student loans help or hurt? *Research in Higher Education*, 1–29.

Hillman, N. W. (2014). College on credit: A multilevel analysis of student loan default. *The Review of Higher Education*, *37*(2), 169-195.

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Hossler, D., Braxton, J. & Coopersmith, G. (1989). Understanding student college choice. *Higher Education: Handbook of Theory and Research*, *5*, 231–288.

Hossler, D. & Gallagher, K. (1987). Studying student college choice: A three-phase model and the implication...-SuperSearch powered by summon. *College and University*, *62*, 201–221.

Hossler, D., Schmit, J. & Vesper, N. (1999). *Going to college: How social, economic, and educational factors influence the decisions students make*. JHU Press.

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., ... & Dilig, R. (2020). The Condition of Education 2020. NCES 2020-144. *National Center for Education Statistics*.

Innes, M., Saba, E., Fischer, K., Gandhi, D., Rudilosso, M. C., Joy, N. M., Karmali, T., Pal, A. & Shah, V. (2018). Fashionable modelling with flux. *CoRR*, *abs/1811.01457*. https://arxiv.org/abs/1811.01457

John, E. P. S. (1990). Price response in enrollment decisions: An analysis of the high school and beyond sophomore cohort. *Research in Higher Education*, *31*(2), 161–176.

Johnson, I. Y. (2019). Destinations of admitted out-of-state students: A case of one institution. *Research in Higher Education*, *60*(3), 315–337.

Johnson, S. (2019). 'No exceptions, no questions asked': Progressives propose legislation canceling all student-loan debt. *The Chronical of Higher Education*, *65*(36).

Jones-White, D. R., Radcliffe, P. M., Lorenz, L. M. & Soria, K. M. (2014). Priced out? *Research in Higher Education*, *55*(4), 329–350.

Kim, J. (2012). Exploring the relationship between state financial aid policy and postsecondary enrollment choices: A focus on income and race differences. *Research in Higher Education*, *53*(2), 123–151.

Kohler, M. & Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, *49*(4), 2231–2249.

Lee, J. (2018). Does merit-based aid promote degree attainment?. *KEDI Journal of Educational Policy*, *15*(1).

Leslie, L. L. & Brinkman, P. T. (1988). *The economic value of higher education. American council on education/macmillan series on higher education.* ERIC.

Letkiewicz, J., Lim, H., Heckman, S., Bartholomae, S., Fox, J. J. & Montalto, C. P. (2014). The path to graduation: Factors predicting on-time graduation rates. *Journal of College Student Retention: Research, Theory & Practice*, *16*(3), 351–371.

Lin, C.-H., Chen, J.-H. & Borden, V. M. (2020). Making graduation matter: Initial evidence and implications for policy and research. *Studies in Higher Education*, 1–16.

Litten, L. H. et al. (1983). *Applying market research in college admissions.*

Loan Debt and Earning by Field of Study. (2019, November) U.S. Department of Education, Retrieved December 12, 2020 from https://collegescorecard.ed.gov/data//

Loans for Undergraduate Students. (2020, May). NCES Retrieved June 25, 2020, from https://nces.ed.gov/programs/coe/indicator_cub.asp

Long, B. T. (2007). The contributions of economics to the study of college access and success. *Teachers College Record*, *109*(10), 2367–2443.

Lubin, M., Dowson, O., Garcia, J. D., Huchette, J., Legat, B. & Vielma, J. P. (2023). JuMP 1.0: Recent improvements to a modeling language for mathematical optimization. *Mathematical Programming Computation*. https://doi.org/10.1007/s12532-023-00239-3

Luna-Torres, M., McKinney, L., Horn, C. L., & Jones, S. (2018). Understanding loan use and debt burden among low-income and minority students at a large urban community college. *Journal of Student Financial Aid*, *48*(1), 1-28.

Maldonado, S., Armelini, G. & Guevara, C. A. (2017). Assessing university enrollment and admission efforts via hierarchical classification and feature selection. *Intelligent Data Analysis*, *21*(4), 945–962.

McElreath, R. (2015). *Statistical rethinking: A bayesian course with examples in r and stan* (p. 364). Chapman; Hall/CRC.

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman; Hall/CRC.

McPherson, M. S. & Schapiro, M. O. (2010). *Keeping college affordable: Government and educational opportunity*. Brookings Institution Press.

Middaugh, M. F., Graham, R. & Shahid, A. (2003). *A study of higher education instructional expenditures: The delaware study of instructional costs and productivity. Research and development report.*

Moline, A. (1987). The relationship of financial aid to student persistence in a commuter institution: A test of a causal model. *Fourth Annual NASSGP/NCHELP Research Network Conference, St. Louis, MO*.

NCES. (2022). *Undergraduate retention and graduation rates*. https://nces.ed.gov/programs/coe/indicator/ctr

Noopila, M. Y., & Williams Pichon, H. (2020). The Influence of Stafford Load Debtload on Persistence Among Historically Underserved Populations at a Hispanic Serving Institution. *Journal of Student Financial Aid*, *49*(3), 3.

Nora, A. & Cabrera, A. F. (1996). The role of perceptions of prejudice and discrimination on the adjustment of minority students to college. *The Journal of Higher Education*, *67*(2), 119–148.

Patel, V. (2020). Look who's talking about canceling debt: How a fringe idea went mainstream. *The Chronical of Higher Education*, *66*(19).

Paulsen, M. B. (1990). *College choice: Understanding student enrollment behavior. ASHE-ERIC higher education report no. 6.* ERIC.

Portes, A. & Wilson, K. L. (1976). Black-white differences in educational attainment. *American Sociological Review*, 414–431.

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

Raisman, N. (2013). The cost of college attrition at four-year colleges & universities- an analysis of 1669 US institutions. *Policy Perspectives*.

Ranzato, M. A., Poultney, C., Chopra, S., & Cun, Y. (2006). Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, *19*.

Sandoval-Palis, I., Naranjo, D., Vidal, J. & Gilar-Corbi, R. (2020). Early dropout prediction model: A case study of university leveling course students. *Sustainability*, *12*(22), 9314.

Schmidt-Hieber, J. (2020). *Nonparametric regression using deep neural networks with ReLU activation function*.

Schneider, M. (2010). Finishing the first lap: The cost of first year student attrition in america's four year colleges and universities. *American Institutes for Research*.

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, *1*(1), 64–85.

St John, E. P. (1989). The influence of student aid on persistence. *Journal of Student Financial Aid*, *19*(3), 5.

St John, E. P. (1991). The impact of student financial aid: A review of recent research. *Journal of Student Financial Aid*, *21*(1), 2.

St John, E. P. & Noell, J. (1987). Student loans and higher education opportunities: Evidence on access, persistence, and choice of major. *Fourth Annual NASSGP/NCHELP Research Network Conference. St. Louis, MO: Washington University*.

Stage, F. K. (1989). Motivation, academic and social integration, and the early dropout. *American Educational Research Journal*, *26*(3), 385–402.

Stage, F. K. & Hossler, D. (1989). Differences in family influences on college attendance plans for male and female ninth graders. *Research in Higher Education*, *30*(3), 301–315.

Stinebrickner, R. & Stinebrickner, T. (2014). Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model. *Journal of Labor Economics*, *32*(3), 601–644.

Sultana, S., Khan, S. & Abbas, M. A. (2017). Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts. *International Journal of Electrical Engineering Education*, *54*(2), 105–118.

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

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

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition.* ERIC.

Tinto, V. (1988). Stages of student departure: Reflections on the longitudinal character of student leaving. *The Journal of Higher Education*, *59*(4), 438–455.

Trends in College Pricing and Student Aid 2020. (2020, October) College Board, Retrieved May 8, 2021, from https://research.collegeboard.org/trends/college-pricing

Tuttle, R. (1981). *A path analytic model of the college going decision.*

USAFacts. (2019, November 1). *Student Debt Explained: Breaking Down the $1.6T in Loans*. U.S. News. Retrieved December 12, 2019, from https://www.usnews.com/news/elections/articles/2019- 11-01/student-debt-explained-breaking-down-the-16t-in-loans

US News, (2019). https://www.usnews.com/news/elections/articles/2019-11-01/student-debt-explained-breaking-down-the-16t-in-loans

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

Venti, S. F. & Wise, D. A. (1983). Individual attributes and self-selection of higher education: College attendance versus college completion. *Journal of Public Economics*, *21*(1), 1–32.

Wagner, K., Volkening, H., Basyigit, S., Merceron, A., Sauer, P. & Pinkwart, N. (2023). Which approach best predicts dropouts in higher education? *In Proceedings of the 15th International Conference on Computer Supported Education (CSEDU 2023)*, *2*, 15–26.

Willis, R. J. & Rosen, S. (1979). Education and self-selection. *Journal of Political Economy*, *87*(5, Part 2), S7–S36.

Witteveen, D., & Attewell, P. (2019). Social dimensions of student debt: A data mining analysis. *Journal of Student Financial Aid*, *49*(1), 1.

Zhan, M., Xiang, X. & Elliott III, W. (2018). How much is too much: Educational loans and college graduation. *Educational Policy*, *32*(7), 993–1017.

Zhang, S., Du, Z., Zhang, L., Lan, H., Liu, S., Li, L., ... & Chen, Y. (2016, October). Cambricon-X: An accelerator for sparse neural networks. In *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)* (pp. 1-12). IEEE.

Appendix A

CODE TO CONSTRUCT BAYESIAN HIERARCHICAL PIECEWISE EXPONENTIAL MODEL

```
using Random, Distributions
using DynamicHMC, DynamicHMC.Diagnostics
using TransformVariables, LogDensityProblems, LogDensityProblemsAD, TransformedLog
Densities
using Zygote

# Model formulation and posterior distribution construction
function (problem::SurvivalProblem)(θ)
 @unpack y, X = problem   # extract the data
 @unpack β0, β, β_high, σ_h, σ_l  = θ      # extract the parameters

 μ_high = zeros(n_var); σ_0 = 0.1; λ_h = 0.5; λ_l = 0.5

 Xβ = (β0[X] + β[0*n_Period .+ X] .* data_fit.FinAid_Rate.+ β[1*n_Period .+ X] .* data_fit.
Pell_Ind+ β[2*n_Period .+ X] .* data_fit.fed_efc_rate.+ β[3*n_Period .+ X] .* data_fit.home_
distance_std + β[4*n_Period .+ X] .* data_fit.Gender_Ind.+ β[5*n_Period .+ X] .* data_fit.Et
h_ASIAN_Ind + β[6*n_Period .+ X] .* data_fit.Eth_BLACK_Ind.+ β[7*n_Period .+ X] .* dat
a_fit.Eth_HISPA_Ind.+ β[8*n_Period .+ X] .* data_fit.Eth_WHITE_Ind.+ β[9*n_Period .+ X]
 .* data_fit.Eth_Multi_Ind.+ β[10*n_Period .+ X] .* data_fit.Pros_Event_Ind.+ β[11*n_Period
 .+ X] .* data_fit.Admit_Honor_Ind.+ β[12*n_Period .+ X] .* data_fit.Diff_Major_Ind
   .+ β[13*n_Period .+ X] .* data_fit.CampusTour_Ever_Ind.+ β[14*n_Period .+ X] .* data_fi
t.DecisionDay_Ever_Ind.+ β[15*n_Period .+ X] .* data_fit.Delay_Review_Ind)

 w = 1.05
 loglike = sum(w .*y .* Xβ .- exp.(Xβ) .* data_fit.Period_length)
 logpri_β0 = sum(logpdf(MultivariateNormal(β0_MLE, σ_0), β0))
 logpri_high = sum(logpdf(MultivariateNormal(μ_high, σ_h), β_high))
 logpri_σ_h = sum(logpdf(Exponential(λ_h), σ_h))

 μ_β = []
 for i in 1:n_var
  μ_temp = β_high[i]*ones(n_Period)
  if i == 1
    μ_β = μ_temp
    else μ_β = vcat(μ_β, μ_temp)
  end
 end
```

```
  logpri_low = sum(logpdf(MultivariateNormal(μ_β, σ_l), β))
  logpri_σ_l = sum(logpdf(Exponential(λ_l), σ_l))

  loglike + logpri_β0 + logpri_high + logpri_low + logpri_σ_h + logpri_σ_l
end

# Mode fitting
t = as((β0 = as(Array, length(β0_init)), β = as(Array, length(β_init)), β_high = as(Array, lengt
h(β_h_init)), σ_h = asℝ₊, σ_l = asℝ₊))
P = TransformedLogDensity(t, p); ∇P = ADgradient(:Zygote, P);

q₀ = vcat(raw_MLE_para.β0, raw_MLE_para.β_FinAid, raw_MLE_para.β_Pell, raw_MLE_p
ara.β_efc, raw_MLE_para.β_home, raw_MLE_para.β_Gender, raw_MLE_para.β_ASIAN, ra
w_MLE_para.β_BLACK, raw_MLE_para.β_HISPA, raw_MLE_para.β_WHITE, raw_MLE_
para.β_Multi, raw_MLE_para.β_Pros_Event, raw_MLE_para.β_Admit_Honor, raw_MLE_pa
ra.β_Diff_Major, raw_MLE_para.β_CampusTour, raw_MLE_para.β_DecisionDay, raw_MLE
_para.β_Delay_Review, raw_MLE_high_para.β_FinAid[1], raw_MLE_high_para.β_Pell[1], r
aw_MLE_high_para.β_efc[1], raw_MLE_high_para.β_home[1], raw_MLE_high_para.β_Gen
der[1], raw_MLE_high_para.β_ASIAN[1], raw_MLE_high_para.β_BLACK[1], raw_MLE_hi
gh_para.β_HISPA[1], raw_MLE_high_para.β_WHITE[1], raw_MLE_high_para.β_Multi[1], r
aw_MLE_high_para.β_Pros_Event[1], raw_MLE_high_para.β_Admit_Honor[1], raw_MLE_
high_para.β_Diff_Major[1], raw_MLE_high_para.β_CampusTour[1], raw_MLE_high_para.β
_DecisionDay[1], raw_MLE_high_para.β_Delay_Review[1], log(1.0), log(0.2) )

results = mcmc_with_warmup(Random.GLOBAL_RNG, ∇P, 1000; initialization = (q=q₀, ) )
```

Appendix B

DIAGRAMS OF PEM, PEM-NN AND PEM-SNN-2

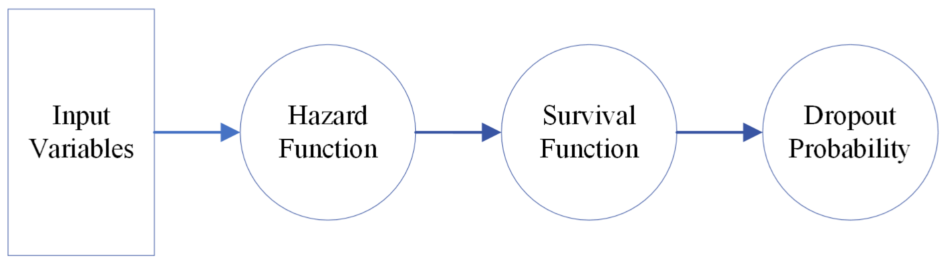

Figure B.1 A diagram of a Piecewise Exponential Model (PEM)

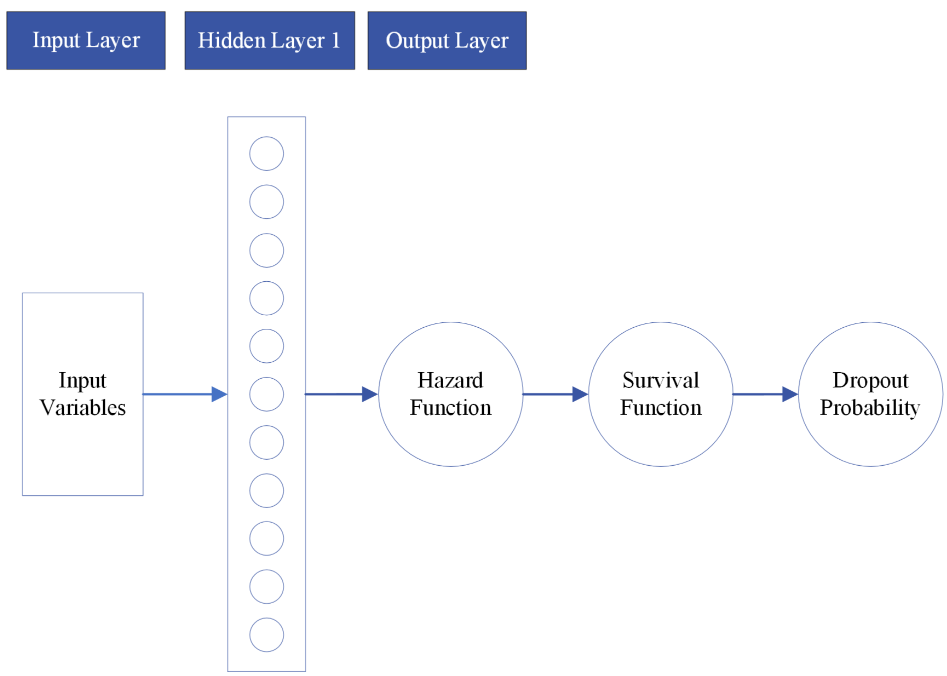

Figure B.2 A diagram of a hybrid model of a fully-connected neural network and a piecewise exponential model (PEM-NN)

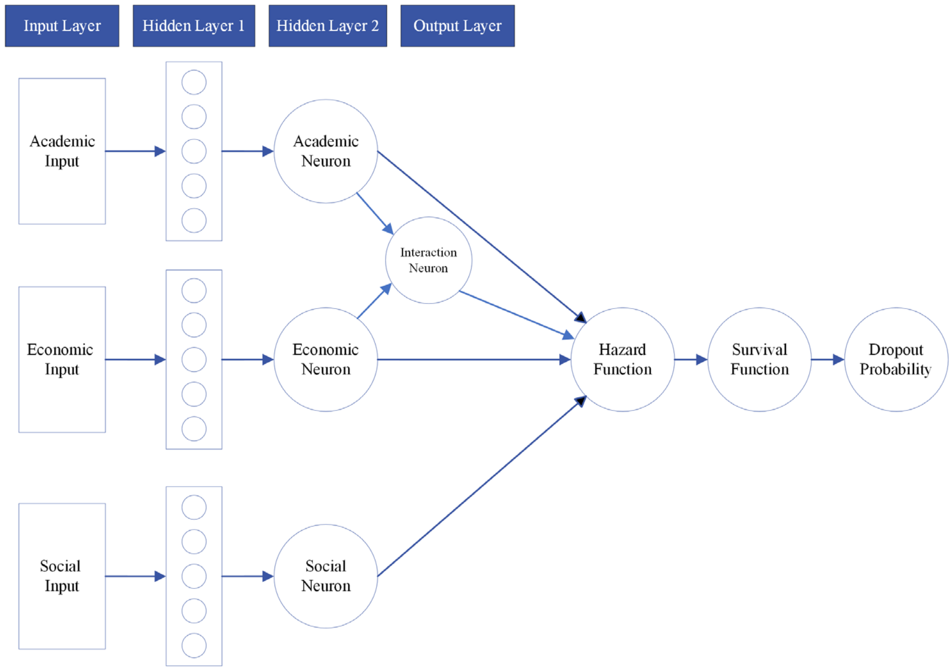

Figure B.3 A diagram of a hybrid model of a structural neural network and a piecewise exponential model, with an interaction between academic and economic integrations (PEM-SNN-2)

Appendix C

CODE TO CONSTRUCT HYBRID MODELS OF PIECEWISE EXPONENTIAL MODELS AND STRUCTURAL NEURAL NETWORKS

```
using Flux
using ProgressMeter
using Distributions


# Model formulation
NN_list = []; n_model = 100

for model_num in 1:n_model

Join(combine, paths) = Parallel(combine, paths)
Join(combine, paths...) = Join(combine, paths)

node_num_1 = 13

model = Chain(Join(hcat,
     Chain(Join(vcat,
     Chain(Dense(n_Acad_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Fin_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Socio_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast))
    ), Dense(3 => 1, exp)),
     Chain(Join(vcat,
     Chain(Dense(n_Acad_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Fin_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Socio_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast))
    ), Dense(3 => 1, exp)),
     Chain(Join(vcat,
     Chain(Dense(n_Acad_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Fin_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Socio_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast))
    ), Dense(3 => 1, exp)),
     Chain(Join(vcat,
     Chain(Dense(n_Acad_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Fin_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Socio_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast))
    ), Dense(3 => 1, exp)),
     Chain(Join(vcat,
     Chain(Dense(n_Acad_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Fin_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Socio_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast))
    ), Dense(3 => 1, exp)),
     Chain(Join(vcat,
     Chain(Dense(n_Acad_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Fin_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast)),
     Chain(Dense(n_Socio_Var => node_num_1, sigmoid_fast; init=Flux.zeros32),Dense(node_num_1 => 1, sigmoid_fast))
```

```
        ), Dense(3 => 1, exp))))

# Construct loss function
function my_loss(result, label)
    w_1 = 1.5
    loss_hat = (w_1 .* label .* vec(log.(result)).- vec(result) .*1)
    -sum(loss_hat)/n_fit
end

# Model fitting
opt_state = Flux.setup(Flux.Adam(0.01), model)
losses = Float32[]

@showprogress for epoch in 1:10000
    input = x_surv; label = y_flat

    val, grads = Flux.withgradient(model) do m
      result = m(input)
      my_loss(result, label)
    end

    push!(losses, val)

    # Detect loss of Inf or NaN. Print a warning, and then skip update!
    if !isfinite(val)
      @warn "loss is $val on $epoch epochs" epoch
      continue
    end

    Flux.update!(opt_state, model, grads[1])

  # Stop training when some criterion is reached
  if  (length(losses) > 2) && (abs(losses[length(losses)-1] - losses[length(losses)]) <1e-7)
    println("stopping after $epoch epochs")
    break
  end
end

append!(NN_list, model)
end
```

Appendix D

DESCRIPTIVE STATISTICS OF STUDENT CHARACTERISTICS FROM THE SECOND TO THE FIFTH SPRING SEMESTERS

Table D.1 Descriptive Statistics of Student Characteristics at the End of the Second Spring

	Mean	SD	Min	Max
Cumulative loan debt (in $1,000s)	9.13	12.69	0	80.38
Total SAT score	1291	121	750	1590
In-state Residency	0.32		0	1
Male	0.42		0	1
First Generation College Student	0.12		0	1
Underrepresented minority Student	0.11		0	1
Cumulative credits passed for GPA	60	8.42	0	143
Count of classes with DFW grade	0.29	0.73	0	6
Credits registered in a spring	14.89	1.89	0	22
AGI unknown	0.20		0	1
Low AGI	0.13		0	1
Medium AGI	0.47		0	1
High AGI	0.20		0	1
Total grant aid (in $1,000s)	3.53	7.35	0	50.18
Total scholarship aid (in $1,000s)	5.63	11.38	0	95.49

Table D.2 Descriptive Statistics of Student Characteristics at the End of the Third Spring

	Mean	SD	Min	Max
Cumulative loan debt (in $1,000s)	14.61	19.56	0	123.34
Total SAT score	1291	120	750	1590
In-state Residency	0.32		0	1
Male	0.41		0	1
First Generation College Student	0.12		0	1
Underrepresented minority Student	0.10		0	1
Cumulative credits passed for GPA	91	11.32	12	174
Count of classes with DFW grade	0.22	0.66	0	7
Credits registered in a spring	14.70	2.45	0	22
AGI unknown	0.19		0	1
Low AGI	0.12		0	1
Medium AGI	0.46		0	1
High AGI	0.22		0	1
Total grant aid (in $1,000s)	4.91	10.4	0	73.67
Total scholarship aid (in $1,000s)	8.49	17.27	0	154.92

Table D.3 *Descriptive Statistics of Student Characteristics at the End of the Fourth Spring*

	Mean	SD	Min	Max
Cumulative loan debt (in $1,000s)	19.20	24.21	0	155.79
Total SAT score	1242	127	750	1590
In-state Residency	0.52		0	1
Male	0.57		0	1
First Generation College Student	0.18		0	1
Underrepresented minority Student	0.15		0	1
Cumulative credits passed for GPA	107	17.79	20	203
Count of classes with DFW grade	0.68	1.14	0	9
Credits registered in a spring	13.76	3.28	0	22
AGI unknown	0.22		0	1
Low AGI	0.18		0	1
Medium AGI	0.45		0	1
High AGI	0.15		0	1
Total grant aid (in $1,000s)	8.12	15.88	0	104.4
Total scholarship aid (in $1,000s)	8.16	27.09	0	203.44

Table D.4 *Descriptive Statistics of Student Characteristics at the End of the Fifth Spring*

	Mean	SD	Min	Max
Cumulative loan debt (in $1,000s)	22.68	26.02	0	170.01
Total SAT score	1237	130	910	1580
In-state Residency	0.62		0	1
Male	0.62		0	1
First Generation College Student	0.22		0	1
Underrepresented minority Student	0.17		0	1
Cumulative credits passed for GPA	110	24.36	38	219
Count of classes with DFW grade	0.90	1.28	0	5
Credits registered in a spring	11.13	5.18	0	19
AGI unknown	0.20		0	1
Low AGI	0.25		0	1
Medium AGI	0.41		0	1
High AGI	0.14		0	1
Total grant aid (in $1,000s)	11.74	21.81	0	133.65
Total scholarship aid (in $1,000s)	4.11	11.84	0	108.04

Appendix E

CREDIBLE INTERVALS OF THE ODDS RATIOS FOR THE FACTOR EFFECTS

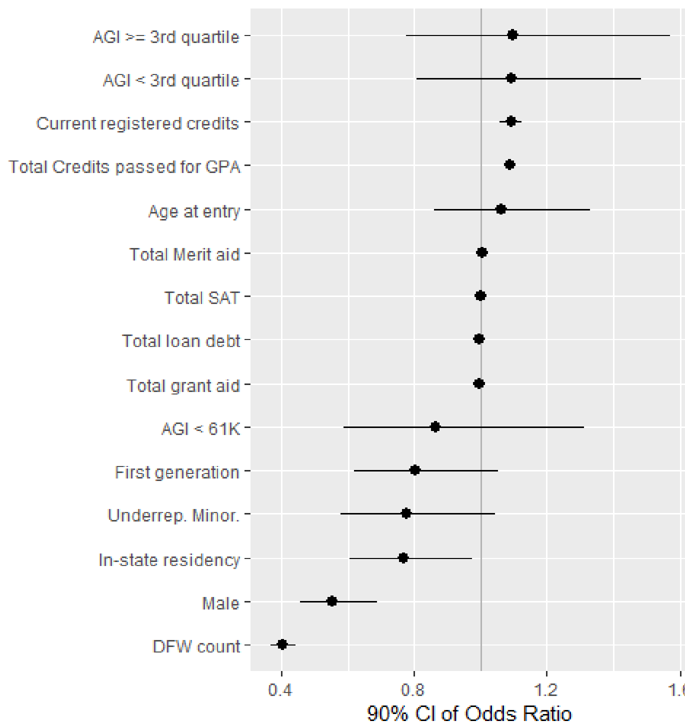

Figure E.1 The 90% Credible Intervals of the Odds Ratios for the Factor Effects on Institutional Level for the Third Spring

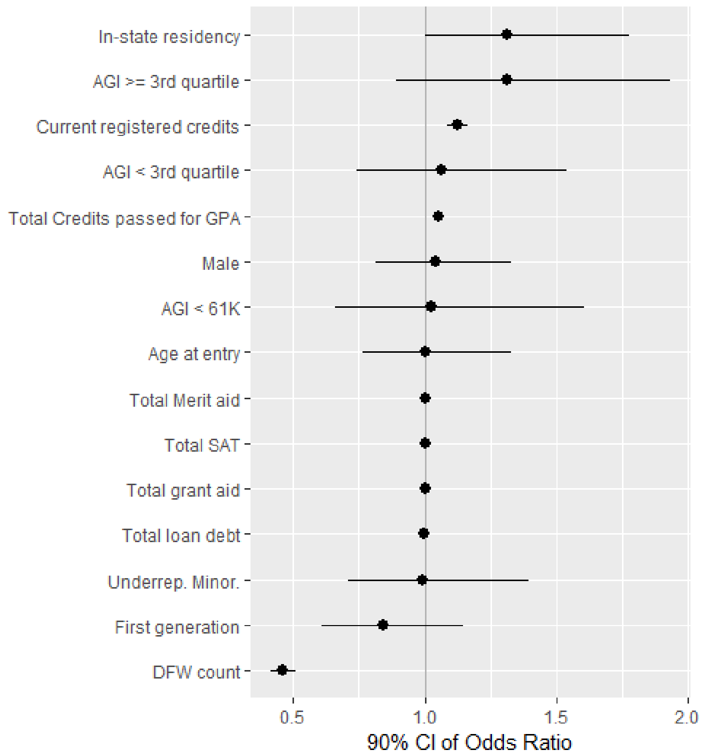

Figure E.2 The 90% Credible Intervals of the Odds Ratios for the Factor Effects on Institutional Level for the Fourth Spring

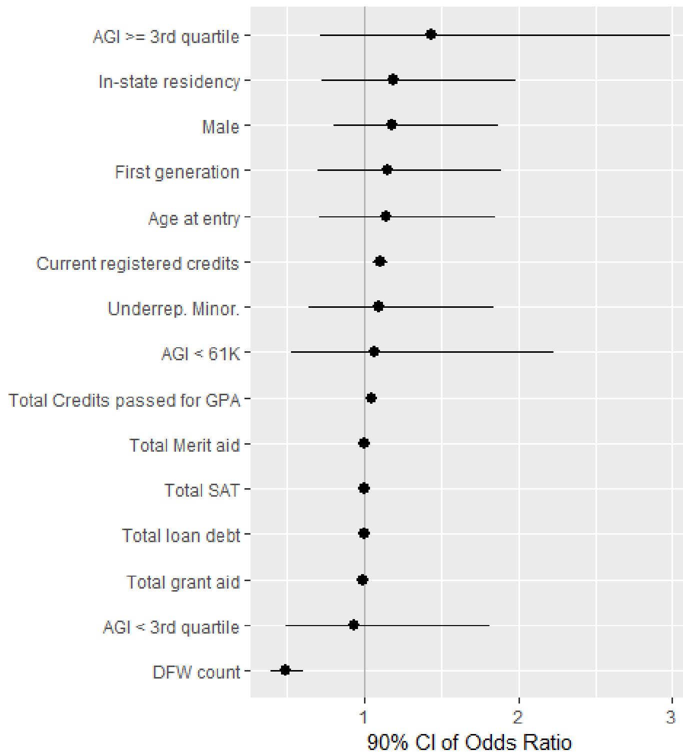

Figure E.3 The 90% Credible Intervals of the Odds Ratios for the Factor Effects on Institutional Level for the Fifth Spring

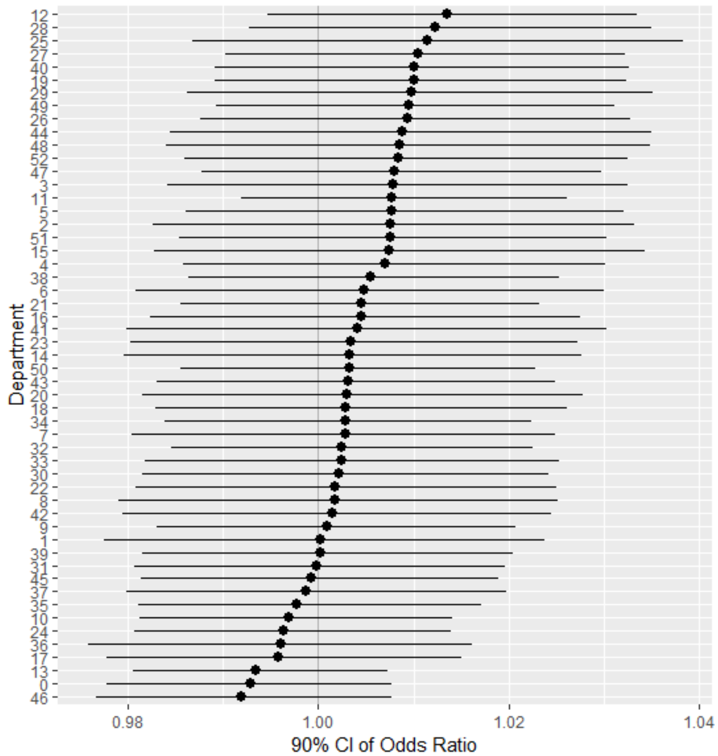

Figure E.4 The 90% Credible Intervals of the Odds Ratios of Loan Debt Effect by Department for the Third Spring

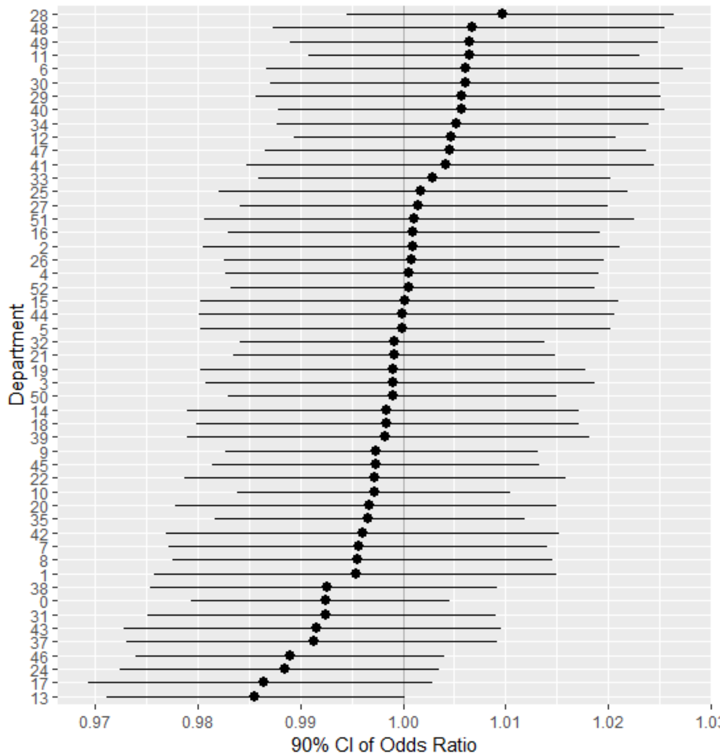

Figure E.5 The 90% Credible Intervals of the Odds Ratios of Loan Debt Effect by Department for the Fourth Spring

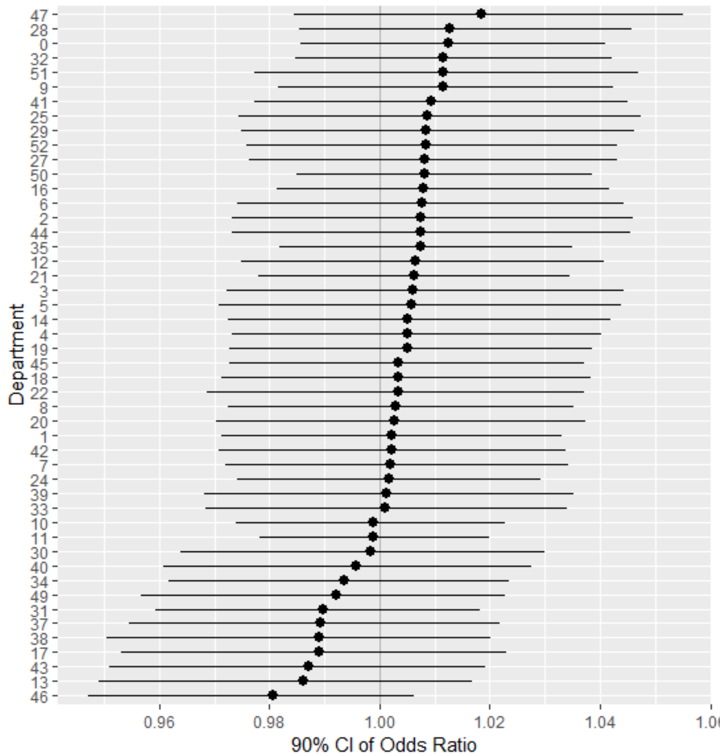

Figure E.6 The 90% Credible Intervals of the Odds Ratios of Loan Debt Effect by Department for the Fifth Spring