USING PREDICTIVE MODELING TO INCREASE SIX-YEAR GRADUATION

Jorge Martinez
Director of Research & Reporting, Enrollment Services

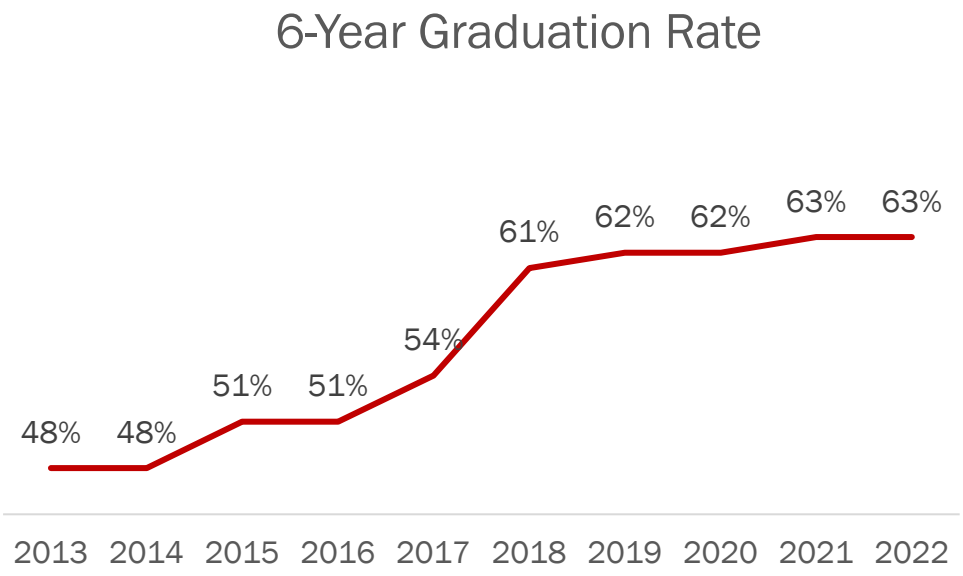
Caroline Neary
Senior Research Analyst, Undergraduate Student Success

UNIVERSITY of HOUSTON
OFFICE OF THE PROVOST

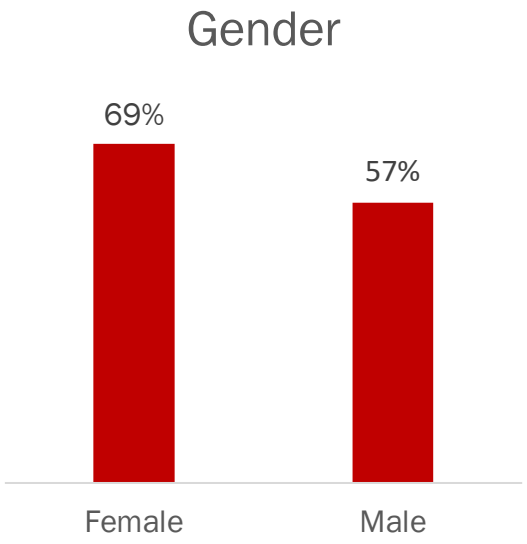
UH: INSTITUTIONAL OVERVIEW

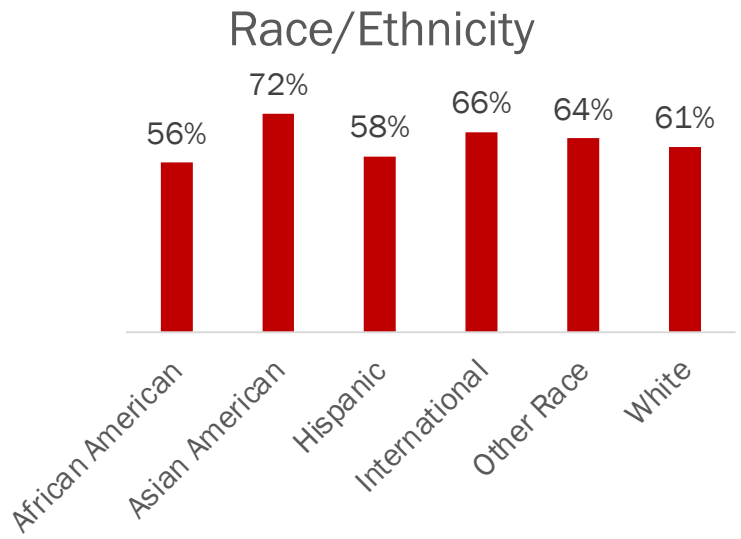
- Large, 4-year, public, urban university
  - 37k undergraduate students
  - 72% attend full-time
  - 45% first-generation
  - 33% Hispanic
  - 40% receive Pell grant

BACKGROUND

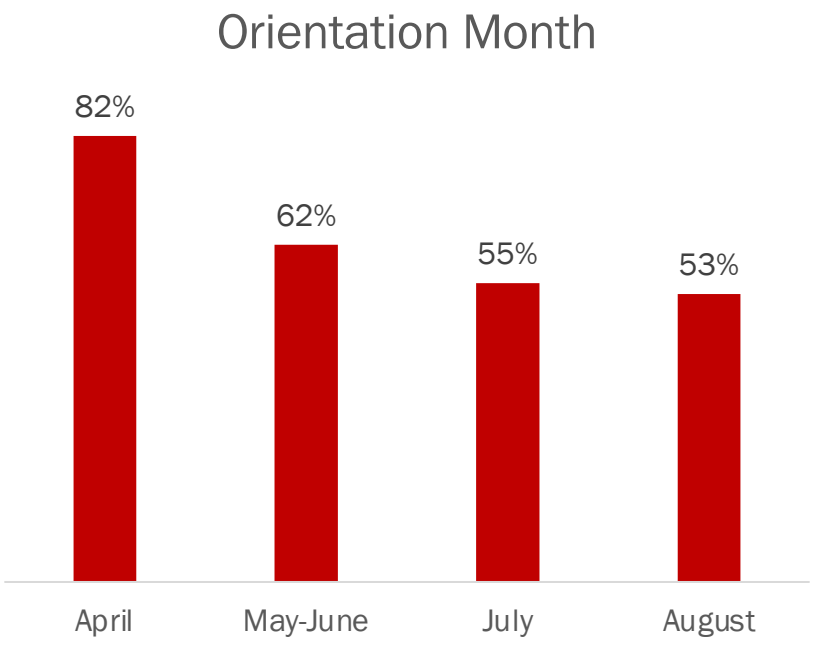

6-Year Graduation Rate
48%
48%
51%
51%
54%
61%
62%
62%
63%
63%
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022

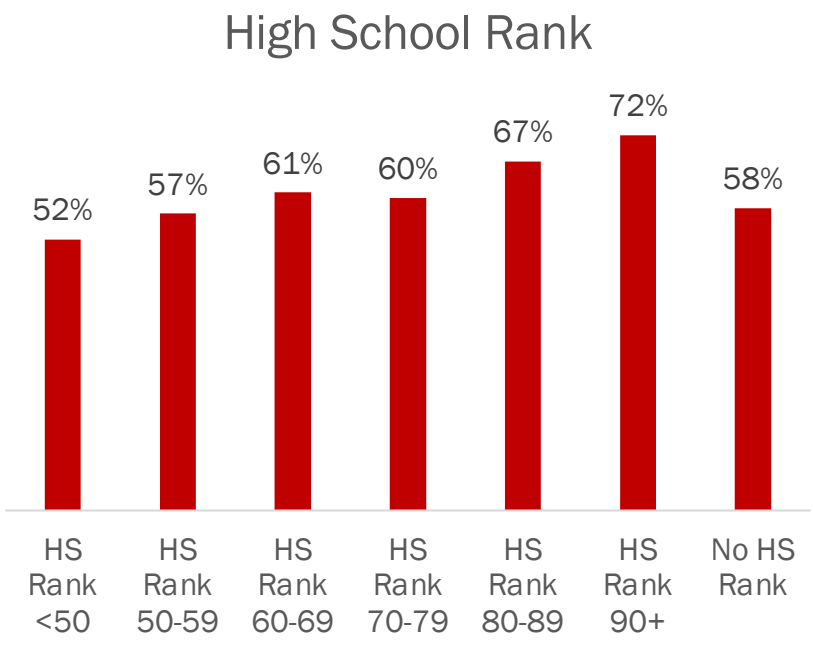
6-YEAR GRADUATION RATES: DEMOGRAPHIC



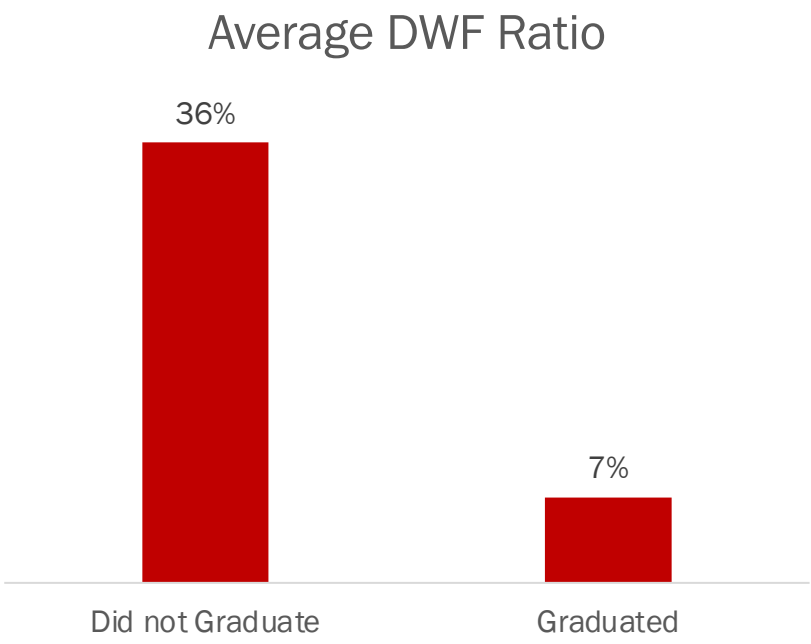


6-YEAR GRADUATION RATES: PRE-COLLEGE

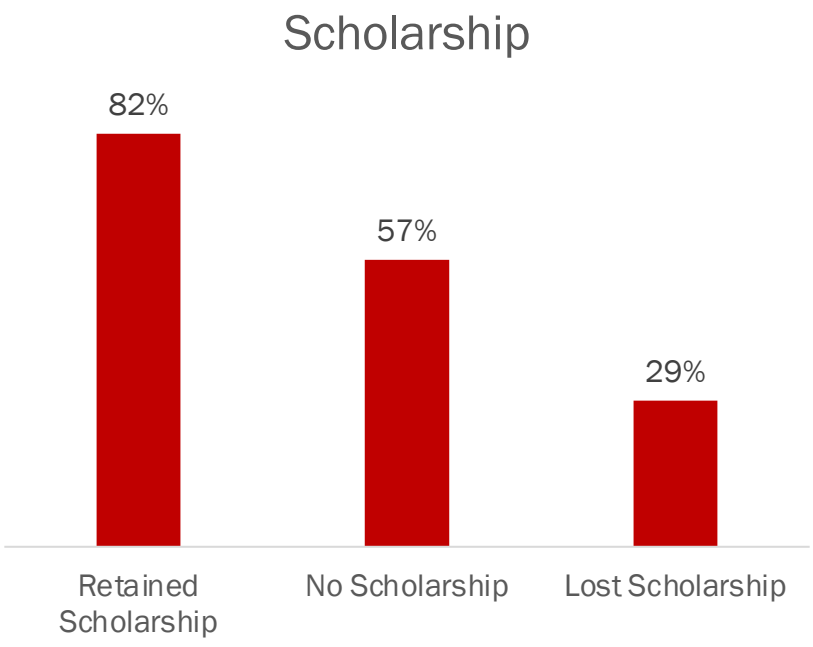


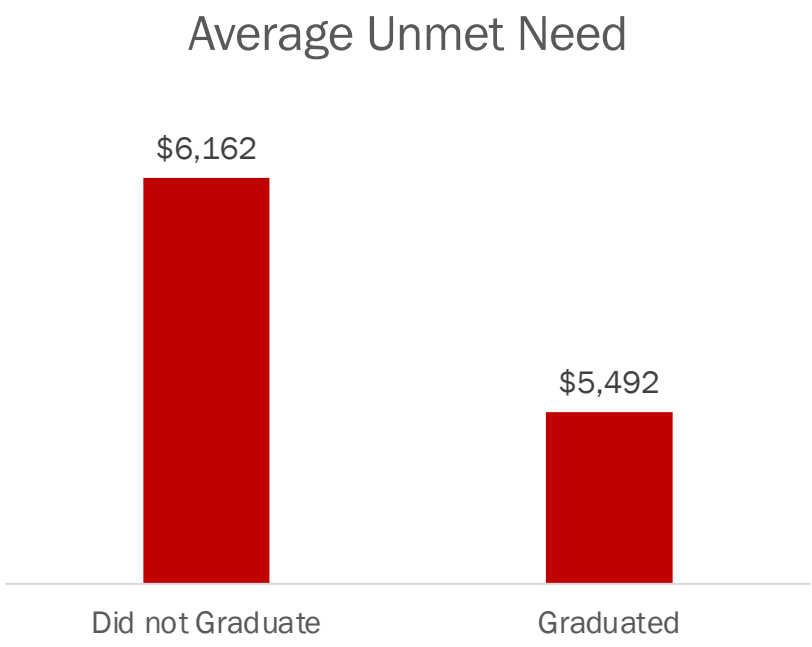


6-YEAR GRADUATION RATES: ACADEMIC



6-YEAR GRADUATION RATES: FINANCIAL





PHASE 1: LOGISTIC REGRESSION MODEL

Objective:

Utilize logistic regression analysis to identify relationships between student characteristics and six-year graduation.

Population:

Fall 2012, Fall 2013, and Fall 2014 FTIC Cohorts (N=10,579)

PREDICTORS

	Academic	Financial	Admissions	Demographics
Strong Predictors	DWF Grade Ratio Total Credits Passed	Lost Scholarship		
Moderate Predictors	Test/Transfer Credits Percent Full-Time Cumulative GPA Change of College	No Scholarship		Race/Ethnicity Residence County
Non-Significant Variables		Pell Eligibility Total Loans Unmet Financial Need	HS Class Rank Orientation Month SAT Score	Gender First Generation

ACTIONABLE CONCLUSIONS

Compared to students from Harris County and its adjacent counties, students from other Texas counties were less likely to graduate in six years.

- ACTION: Support and outreach for these students (about 14% of FA20 cohort)

Students who lost or never had a merit scholarship were less likely to graduate in six years.

- ACTION: Expand first year academic scholarship opportunities, e.g., retention scholarship

Students enrolled full-time for a higher percentage of terms were more likely to graduate in six years.

- ACTION: Continue to encourage full-time enrollment, e.g., UHin4

Students with a higher ratio of D, W, and F grades to all grades were less likely to graduate in six years.

- ACTION: Expand support for students/instructors in high DWF rate courses, e.g., Gateways to Completion, LAUNCH

PHASE II: SURVIVAL ANALYSIS

- Helps us answer questions like
  - How long can we expect patients to survive with certain medical conditions?

PHASE II: SURVIVAL ANALYSIS

- Helps us answer questions like
  - How long can we expect **students** to **graduate** with **different characteristics** (gender, college, first generation status)?
  - What proportion of students are expected to graduate **by a specific academic year**?
  - What variables/factors/interventions are likely to **increase or decrease** time to graduation?

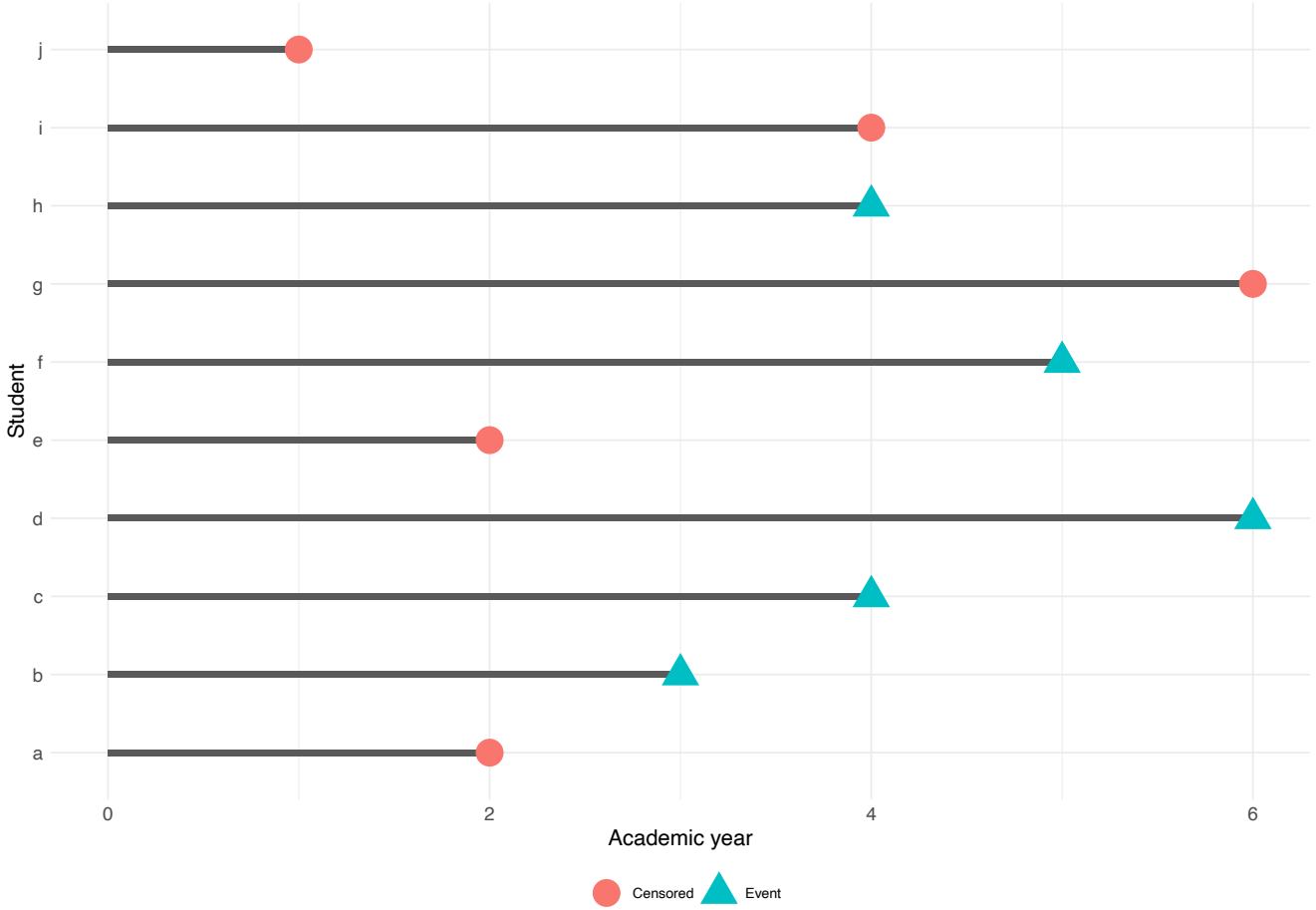
PHASE II: SURVIVAL ANALYSIS

- From classification to **degree velocity**
  - Logistic regression (graduated Y/N)
  - Survival analysis (time-to-degree)
- Model time until an event occurs
  - Compare between groups
  - How event correlates with quantitative variables
- Also known as **Event History Analysis**

CENSORING

- Censoring is a type of **missing data** problem
  - The **event never occurs** during the study window
  - Student **drops out** of the study for various reasons
  - You only know if the individual **survived up to the loss of follow-up**

TIME-TO-EVENT

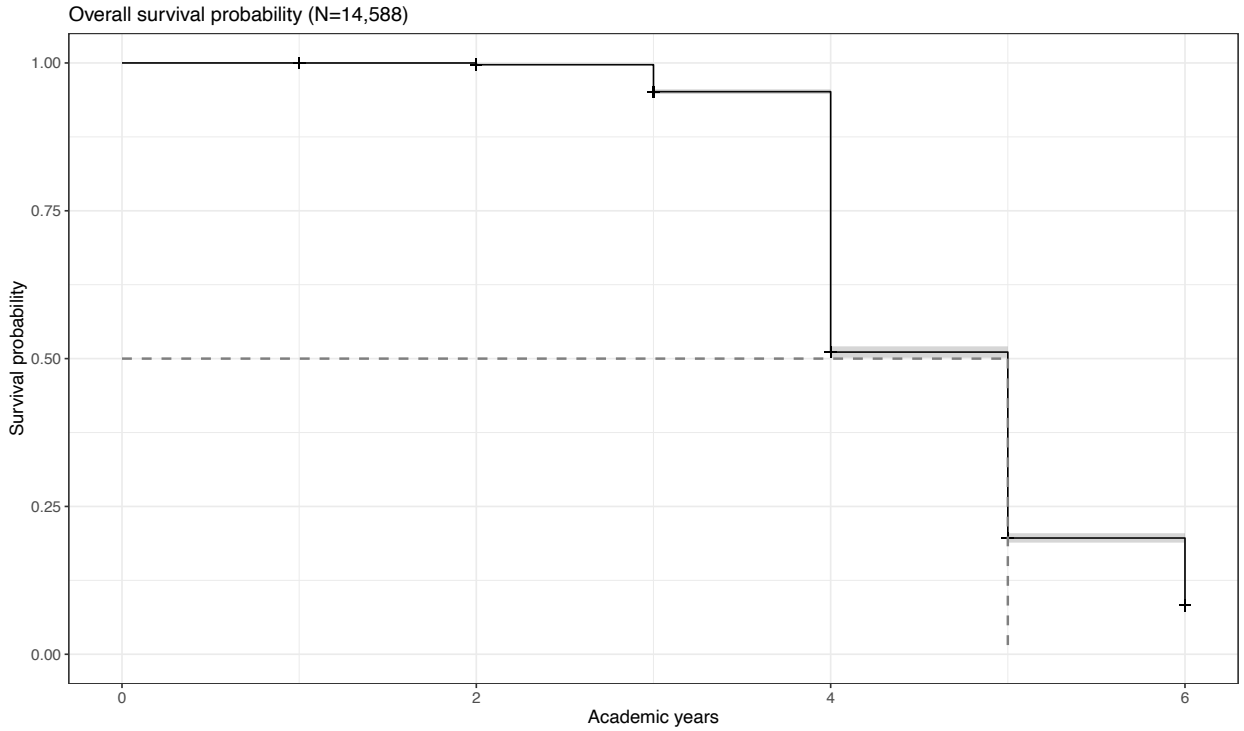


SURVIVAL FUNCTION

- **Survival function** is the probability an individual survives up to and including time *t*.

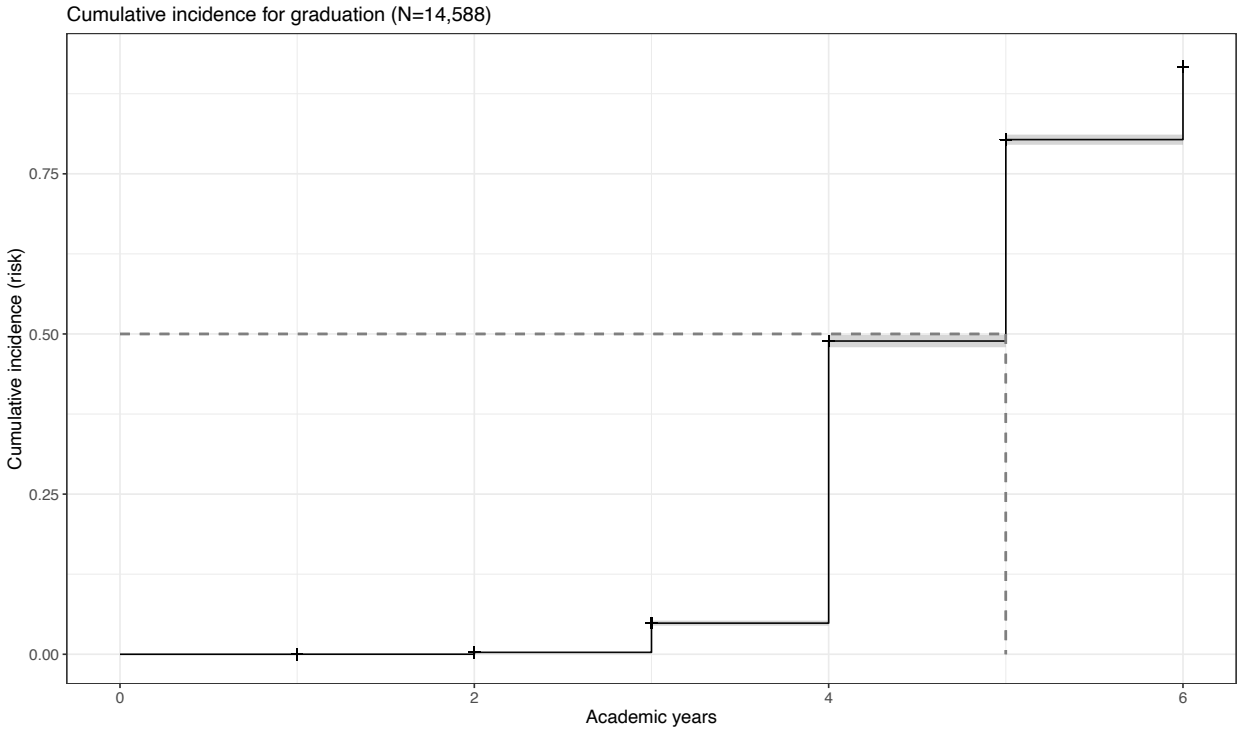
Academic Years	# risk	# event	# censored	Survival probability	Std. err.	Upper	Lower
1	14,588	1	1797	0.9999	0.0001	1.0000	0.9998
2	12,790	37	1607	0.9970	0.0005	0.9980	0.9961
3	11,146	510	817	0.9514	0.0021	0.9554	0.9475
4	9,819	4545	457	0.5110	0.0096	0.5207	0.5015
5	4,817	2964	323	0.1966	0.0206	0.2047	0.1888
6	1,530	887	643	0.0826	0.0364	0.0887	0.0769

KAPLAN-MEIER SURVIVAL CURVE



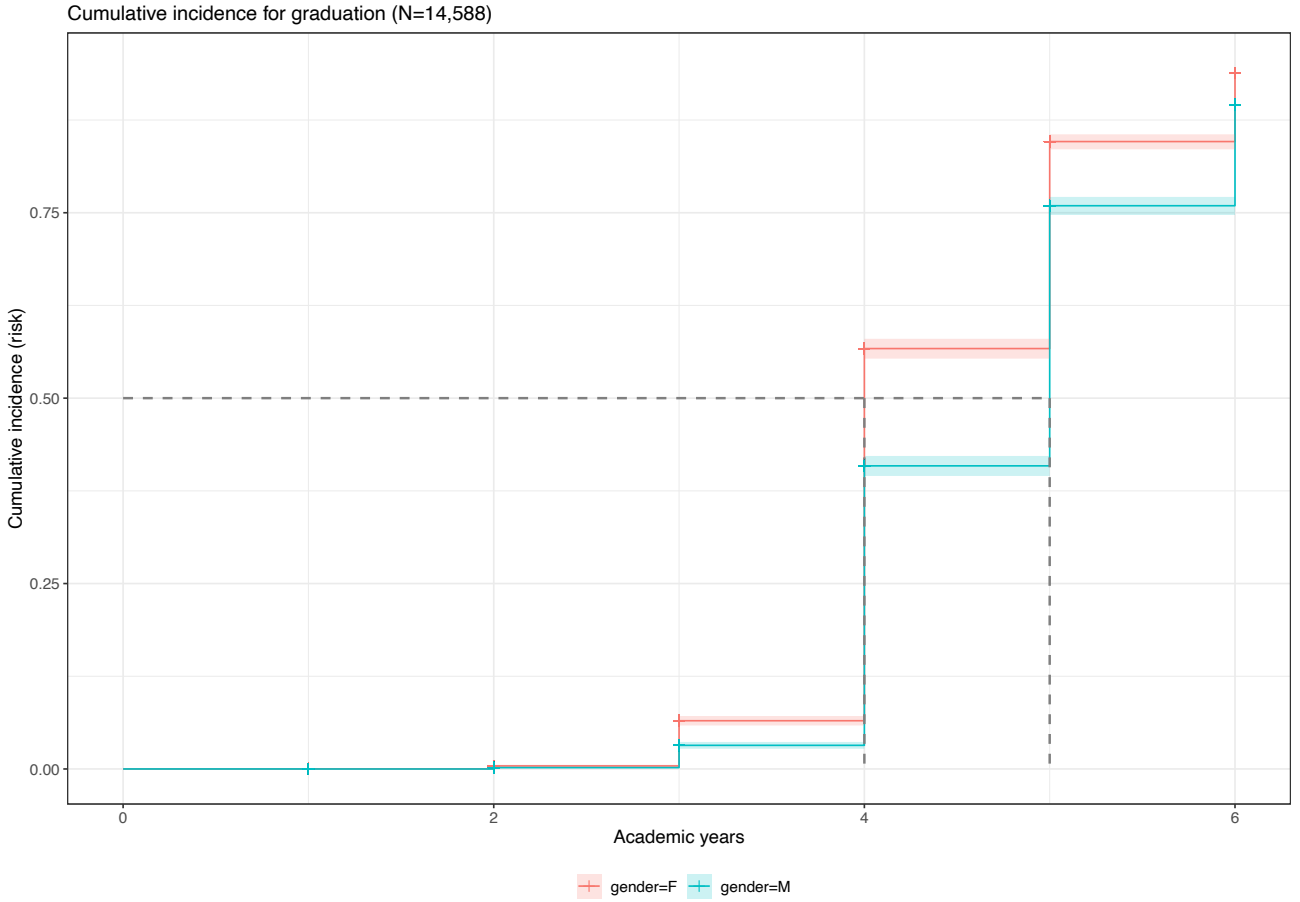
At Risk	14588	12790	9819	1530
Censored	0	3404	4678	5644
Events	0	38	5093	8944

CUMULATIVE INCIDENCE

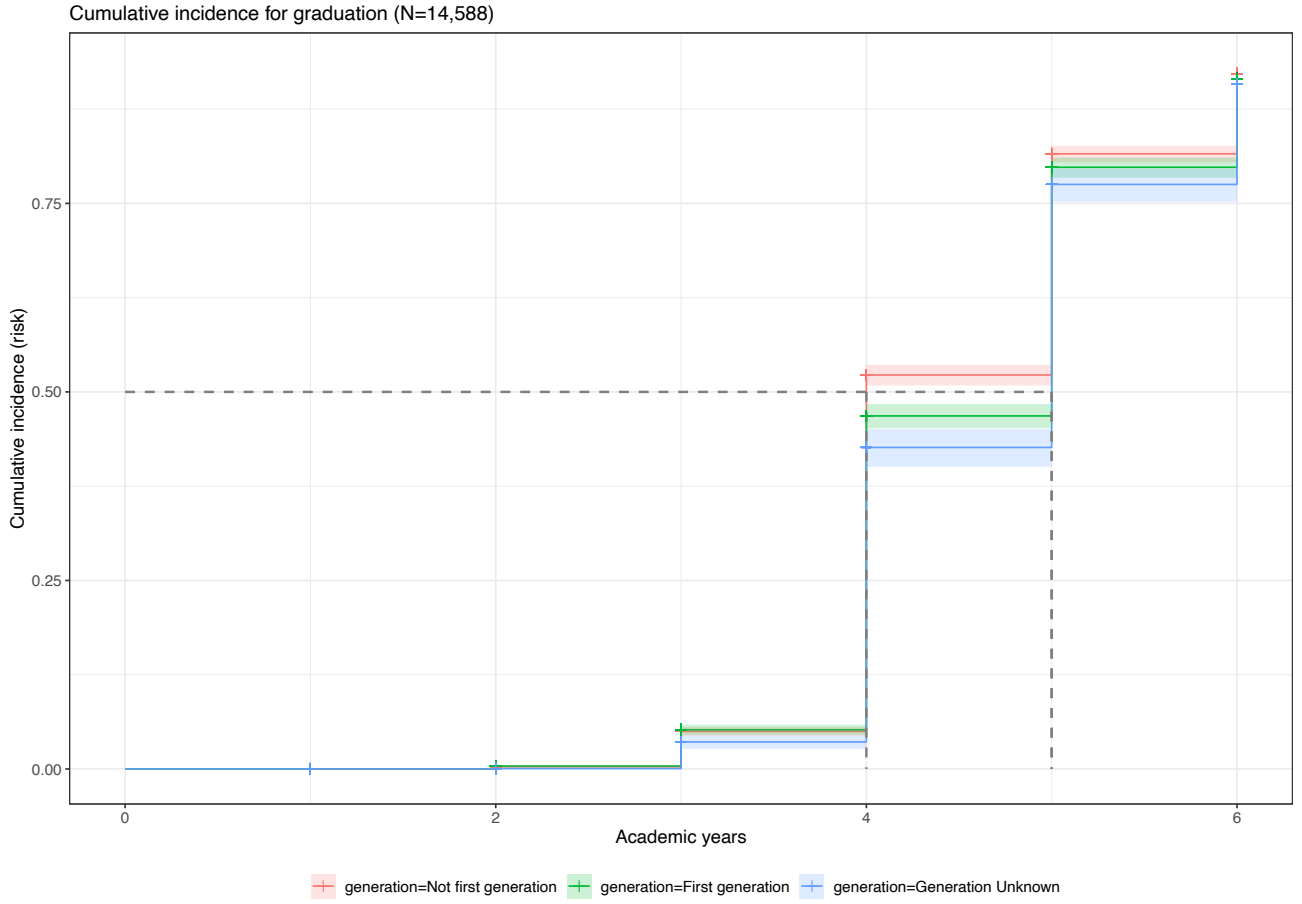


At Risk	14588	12790	9819	1530
Censored	0	3404	4678	5644
Events	0	38	5093	8944

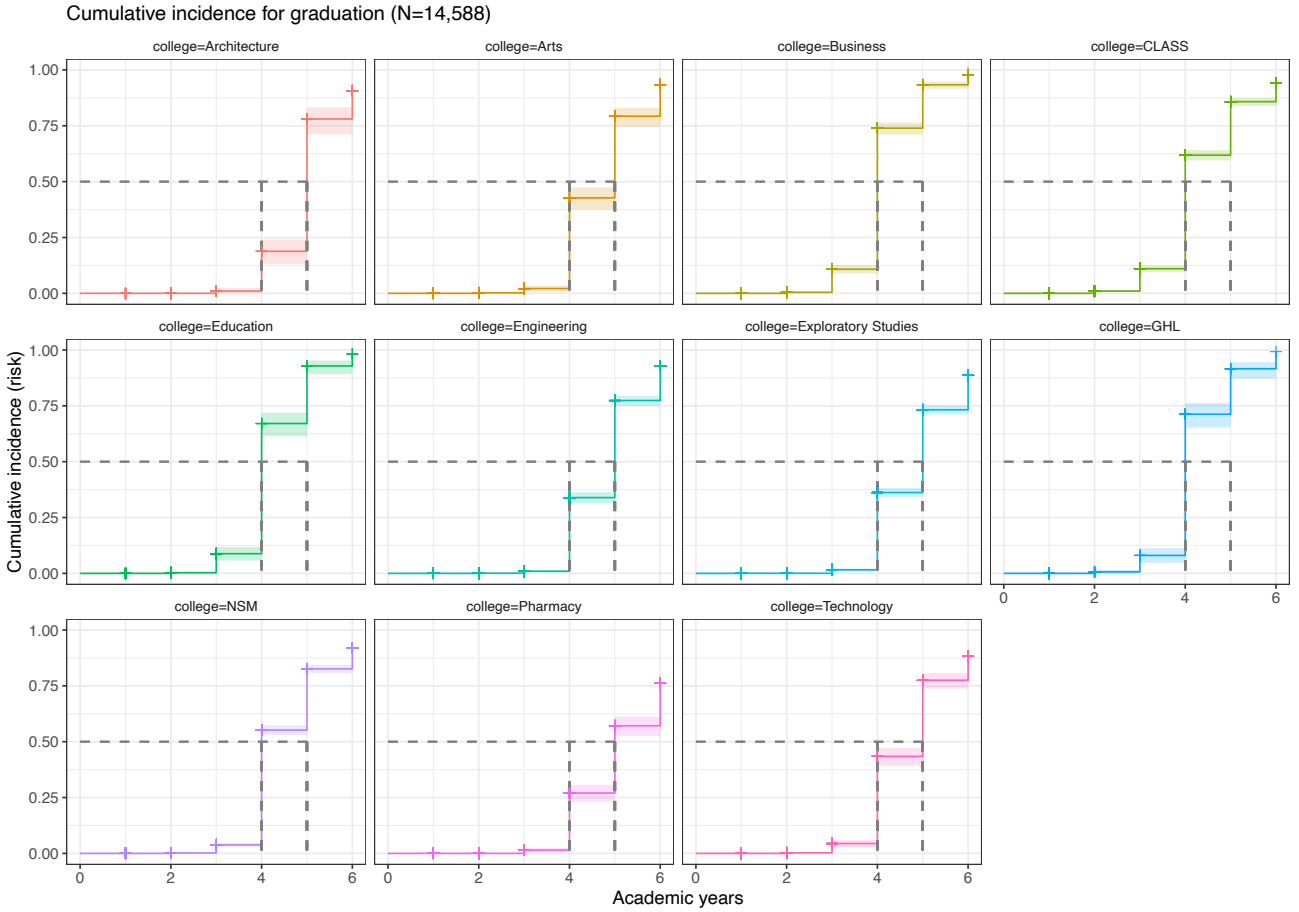
GENDER


Cumulative incidence for graduation (N=14,588)
Cumulative incidence (risk)
0.75
0.50
0.25
0.00
0
2
4
6
Academic years
gender=F
gender=M

GENERATION


Cumulative incidence for graduation (N=14,588)
Cumulative incidence (risk)
0.75
0.50
0.25
0.00
0
2
4
6
Academic years
generation=Not first generation
generation=First generation
generation=Generation Unknown

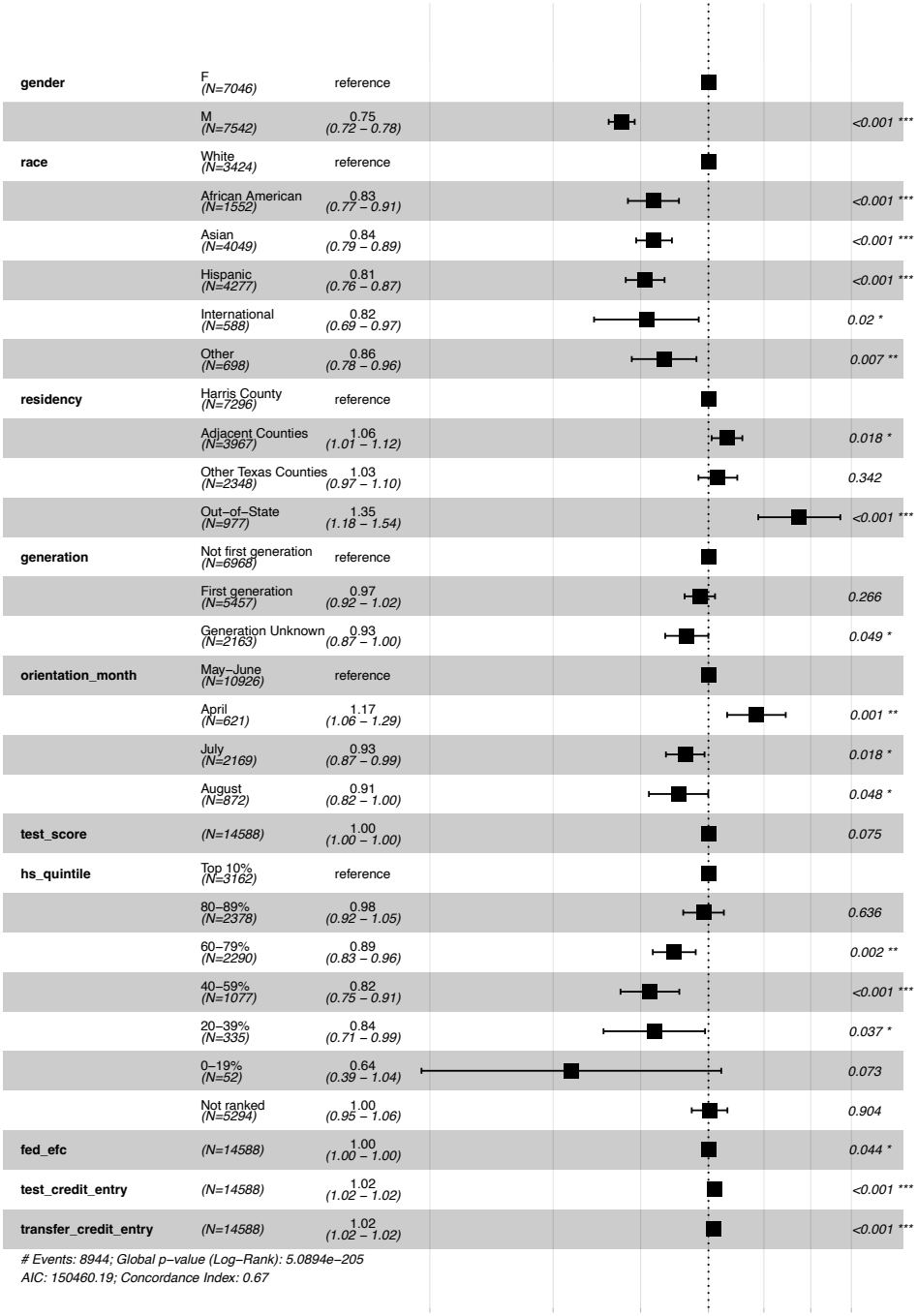
COLLEGE

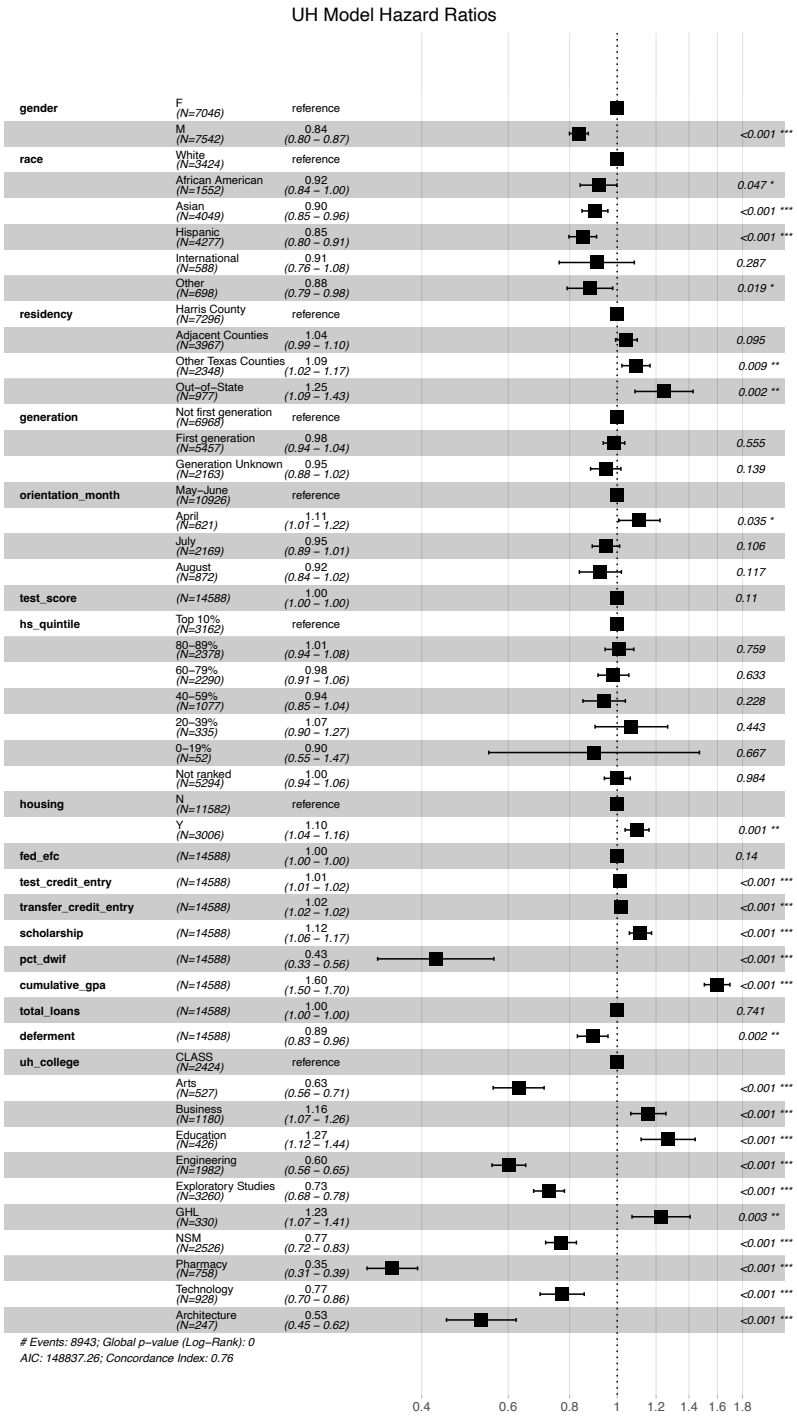

Cumulative incidence for graduation (N=14,588)
college=Architecture
college=Arts
college=Business
college=CLASS
college=Education
college=Engineering
college=Exploratory Studies
college=GHL
college=NSM
college=Pharmacy
college=Technology
Cumulative incidence (risk)
1.00
0.75
0.50
0.25
0.00
0
2
4
6
Academic years

COX PROPORTIONAL HAZARD MODELS

- Survival function helps us **compare rates between categorical** values
- We need the **hazard function** to estimate models with covariates and covariates that are numeric
- The **hazard** is the instantaneous event rate at a particular time point *t*.
- **Hazard ratio** is the ratio of two rates between two levels of a predictor (or unit increase in continuous predictor)

Incoming Characteristics Model Hazard Ratios




UH Model Hazard Ratios
gender
F (N=7046) reference
M (N=7542) 0.84 (0.80 – 0.87) <0.001 ***
race
White (N=3424) reference
African American (N=1552) 0.92 (0.84 – 1.00) 0.047 *
Asian (N=4049) 0.90 (0.85 – 0.96) <0.001 ***
Hispanic (N=4277) 0.85 (0.80 – 0.91) <0.001 ***
International (N=588) 0.91 (0.76 – 1.08) 0.287
Other (N=698) 0.88 (0.79 – 0.98) 0.019 *
residency
Harris County (N=7296) reference
Adjacent Counties (N=3967) 1.04 (0.99 – 1.10) 0.095
Other Texas Counties (N=2348) 1.09 (1.02 – 1.17) 0.009 **
Out–of–State (N=977) 1.25 (1.09 – 1.43) 0.002 **
generation
Not first generation (N=6968) reference
First generation (N=5457) 0.98 (0.94 – 1.04) 0.555
Generation Unknown (N=2163) 0.95 (0.88 – 1.02) 0.139
orientation_month
May–June (N=10926) reference
April (N=621) 1.11 (1.01 – 1.22) 0.035 *
July (N=2169) 0.95 (0.89 – 1.01) 0.106
August (N=872) 0.92 (0.84 – 1.02) 0.117
test_score (N=14588) 1.00 (1.00 – 1.00) 0.11
hs_quintile
Top 10% (N=3162) reference
80–89% (N=2378) 1.01 (0.94 – 1.08) 0.759
60–79% (N=2290) 0.98 (0.91 – 1.06) 0.633
40–59% (N=1077) 0.94 (0.85 – 1.04) 0.228
20–39% (N=335) 1.07 (0.90 – 1.27) 0.443
0–19% (N=52) 0.90 (0.55 – 1.47) 0.667
Not ranked (N=5294) 1.00 (0.94 – 1.06) 0.984
housing
N (N=11582) reference
Y (N=3006) 1.10 (1.04 – 1.16) 0.001 **
fed_efc (N=14588) 1.00 (1.00 – 1.00) 0.14
test_credit_entry (N=14588) 1.01 (1.01 – 1.02) <0.001 ***
transfer_credit_entry (N=14588) 1.02 (1.02 – 1.02) <0.001 ***
scholarship (N=14588) 1.12 (1.06 – 1.17) <0.001 ***
pct_dwif (N=14588) 0.43 (0.33 – 0.56) <0.001 ***
cumulative_gpa (N=14588) 1.60 (1.50 – 1.70) <0.001 ***
total_loans (N=14588) 1.00 (1.00 – 1.00) 0.741
deferment (N=14588) 0.89 (0.83 – 0.96) 0.002 **
uh_college
CLASS (N=2424) reference
Arts (N=527) 0.63 (0.56 – 0.71) <0.001 ***
Business (N=1180) 1.16 (1.07 – 1.26) <0.001 ***
Education (N=426) 1.27 (1.12 – 1.44) <0.001 ***
Engineering (N=1982) 0.60 (0.56 – 0.65) <0.001 ***
Exploratory Studies (N=3260) 0.73 (0.68 – 0.78) <0.001 ***
GHL (N=330) 1.23 (1.07 – 1.41) 0.003 **
NSM (N=2526) 0.77 (0.72 – 0.83) <0.001 ***
Pharmacy (N=758) 0.35 (0.31 – 0.39) <0.001 ***
Technology (N=928) 0.77 (0.70 – 0.86) <0.001 ***
Architecture (N=247) 0.53 (0.45 – 0.62) <0.001 ***
Events: 8943; Global p–value (Log–Rank): 0
AIC: 148837.26; Concordance Index: 0.76
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8

MODEL COMPARISONS

Incoming vs. First Year survival models:

- Pre-college characteristics no longer significant once more college characteristics were incorporated into the first-year model

MODEL COMPARISONS

Logistic Regression vs. Survival Analysis:

- Gender, race/ethnicity = African American, and race/ethnicity = Hispanic became significant in the survival analysis
- Being from further away from UH became significant in the survival analysis with a positive relationship to graduation

NEXT STEPS

- Decide on the most parsimonious model
- Expand analysis term-by-term
- Time-varying covariates
- Incorporate course data
- Use to identify students for outreach/intervention at specific times

LIMITATIONS

The variables in the model are limited to the data accessible on UH students. The model does not capture variables like student engagement or sense of belonging; it cannot capture individual student experiences and struggles. It also does not capture the daily efforts of undergraduate student success staff, such as advising, outreach, and tutoring.

CONTACT INFORMATION

Jorge Martinez
jxm@uh.edu

Caroline Neary
csneary@uh.edu