NCAIR 2024 Annual Conference

Institutional Innovation

Optimization of Markov Chain Modeling in Predicting Student Retention

Kien (Kyle) Nguyen, Office of Institutional Research

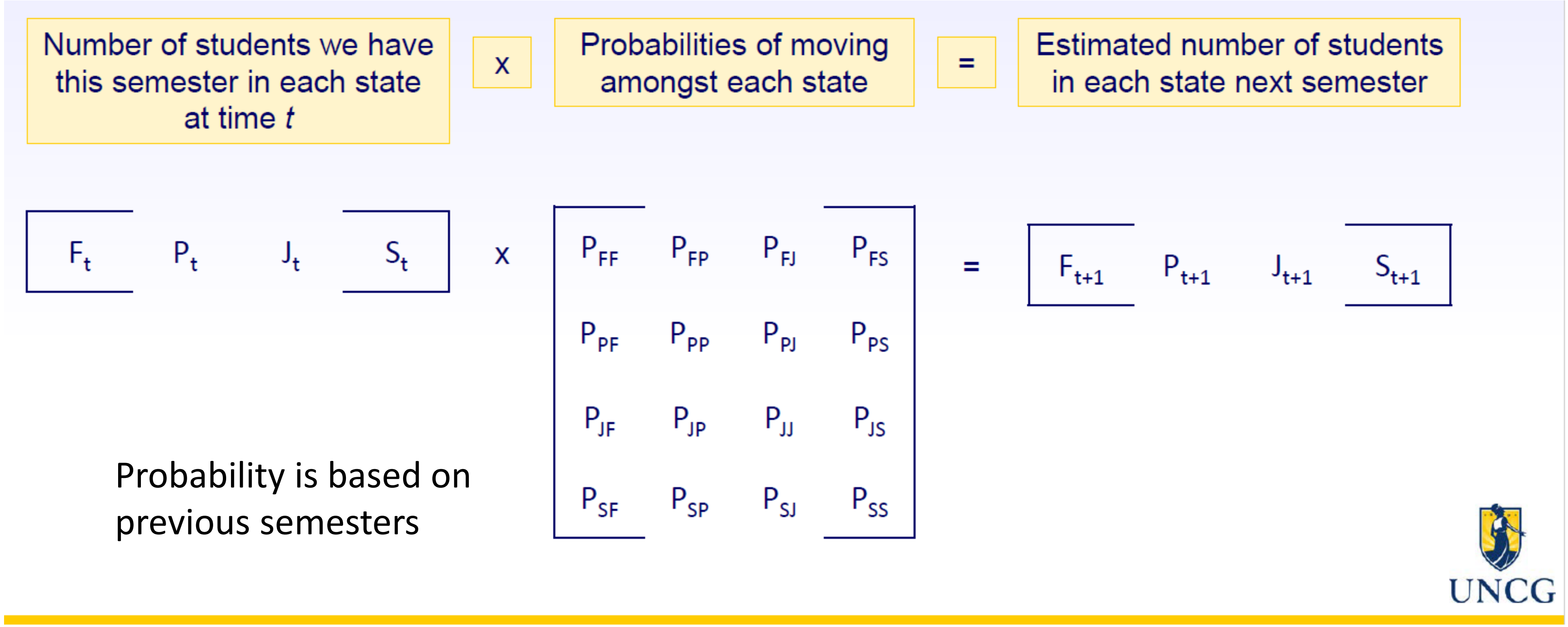
University of North Carolina at Pembroke

Markov Chain Analysis

- Based on Bradley (UNCG, 2018) and Sears (AppState, 2022) presentation about Markov Chain modeling in student retention projection.
- Markov Chain is a stochastic (random) method to describe a process with randomness.
- For each state, there is a fixed number of possible future status. The probability of each status is derived based on past observations (Dai and An, 2018).
- Have been used in higher education in enrollment prediction (Fatima et al. 2022, Gandy et al. 2019), student progression and graduation (Brezavšček et al. 2017, Keener. 2022)
- Reliable, easy to interpret and explain

Markov Chain Analysis

- Methodology: Bradley (2018)



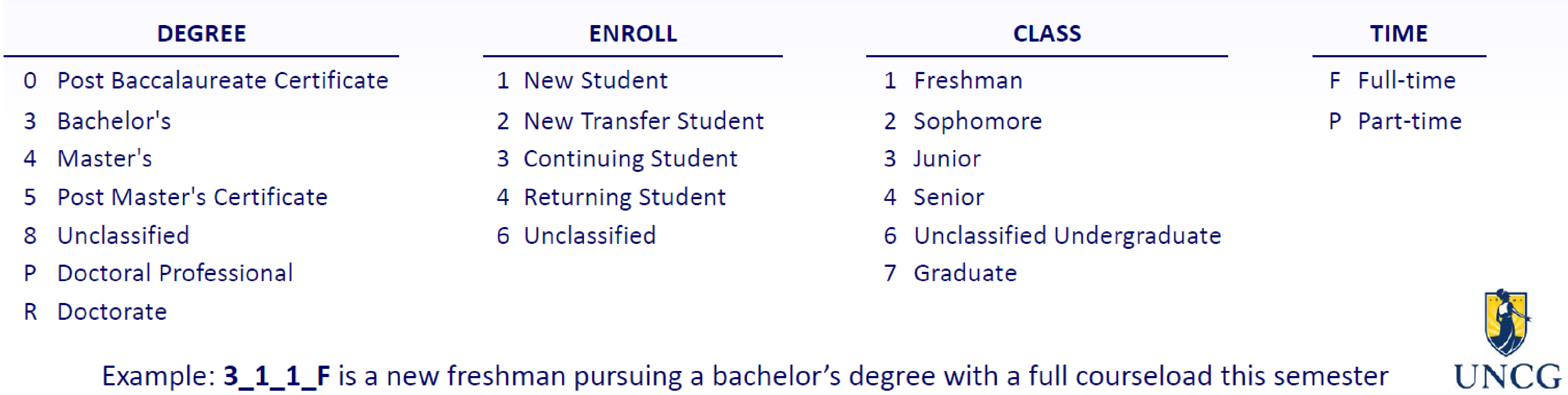
Number of students we have this semester in each state at time *t* x Probabilities of moving amongst each state = Estimated number of students in each state next semester

$$\begin{bmatrix} F_t & P_t & J_t & S_t \end{bmatrix} \times \begin{bmatrix} P_{FF} & P_{FP} & P_{FJ} & P_{FS} \\ P_{PF} & P_{PP} & P_{PJ} & P_{PS} \\ P_{JF} & P_{JP} & P_{JJ} & P_{JS} \\ P_{SF} & P_{SP} & P_{SJ} & P_{SS} \end{bmatrix} = \begin{bmatrix} F_{t+1} & P_{t+1} & J_{t+1} & S_{t+1} \end{bmatrix}$$

Probability is based on previous semesters

Markov Chain Analysis

- Methodology: Bradley (2018)



DEGREE		ENROLL		CLASS		TIME	
0	Post Baccalaureate Certificate	1	New Student	1	Freshman	F	Full-time
3	Bachelor's	2	New Transfer Student	2	Sophomore	P	Part-time
4	Master's	3	Continuing Student	3	Junior		
5	Post Master's Certificate	4	Returning Student	4	Senior		
8	Unclassified	6	Unclassified	6	Unclassified Undergraduate		
P	Doctoral Professional			7	Graduate		
R	Doctorate						

Example: **3_1_1_F** is a new freshman pursuing a bachelor’s degree with a full courseload this semester

UNCG

Markov Chain Analysis

- Determine which parameters to group the student population
- Some common parameters: career, class level, enrollment status
- This is good enough to model progression or time-to-degree
- Goal: To maximize accuracy of retention prediction by choosing optimal set of grouping parameters

Proposed Process

- We use Markov Chain to model next term retention rate
- Binary output: enrolled/not enrolled, excluding graduated and suspended students
- Propose an algorithm to determine list of parameters that yield most accurate results when used as grouping
  - Choose parameters from list
  - Perform Markov Chain process, compare results with actual data
  - Calculate score and compare
  - Choose set of input that yield best results

Proposed Process

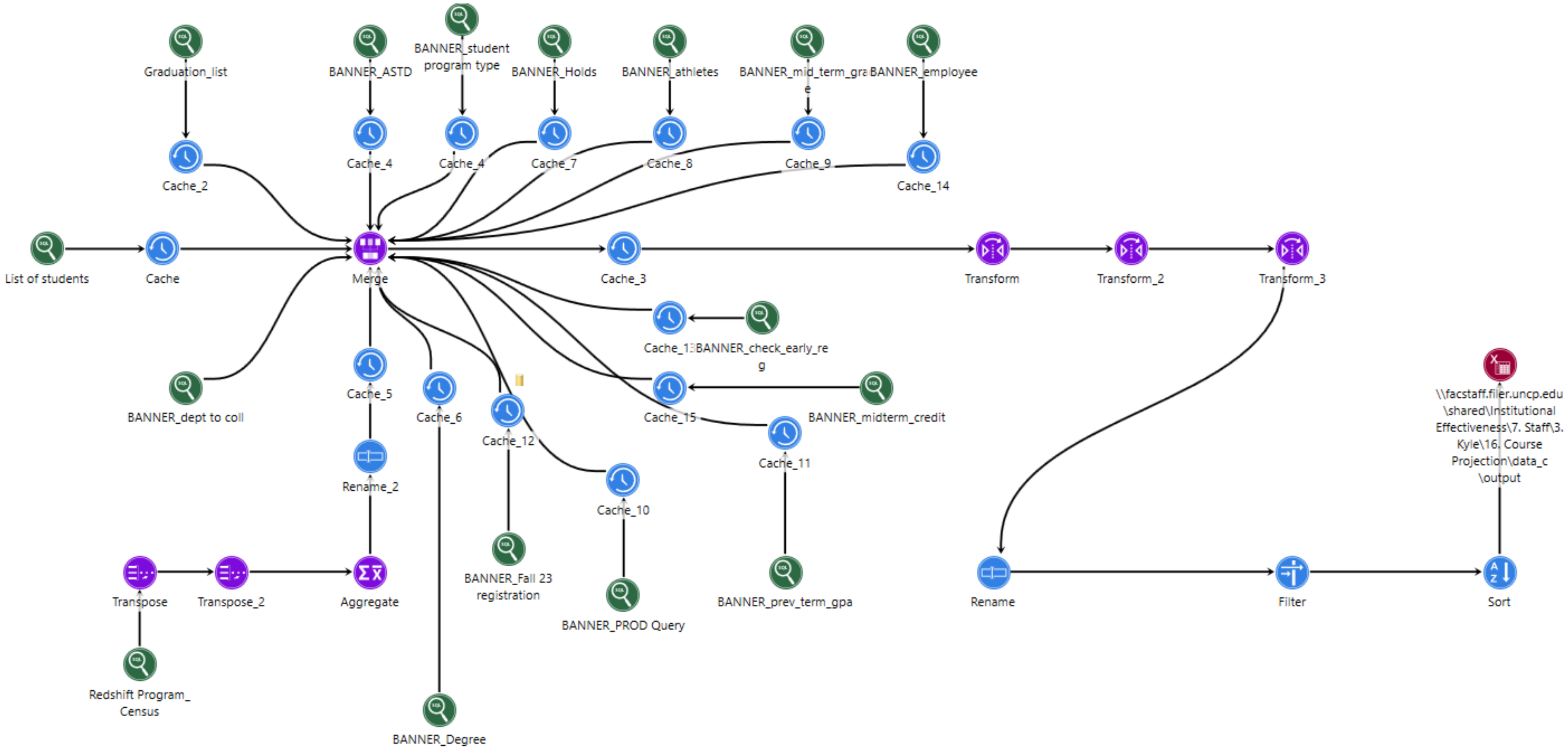
- Markov Chain approach vs machine learning algorithms
- Our goal: This is the initial student population. How many will return next semester overall? How many of freshmen group will return ?..
- Personal experience with papers employing machine learning: focus more on identifying individuals with high probability of not returning.

Algorithm Definition

- List of over 50 parameters from census and live data:
  - career_code
  - class_level
  - pell_status, financial aid status
  - fte_category
  - mid_term_flag
  - Military_affiliated_flag
  - Previous_term_enroll_flag
  - Race, gender, gpa, major, active holds
- Maximum of 8-12 for undergraduate and 6-8 for graduate for UNCP

Algorithm Definition

- Veera workflow to obtain data



Algorithm Definition

- The model divides population into subgroups based on input parameters: career, course load, class level, GPA category…
- Calculate next term retention rate based on most recent two term rate for each subgroup

Group	Fall 2022 retention rate	Fall 2021 retention rate	Fall 2023 predicted retention rate
Full time, freshman, 3-3.5 GPA, Pell eligible, in-state	85%	81%	• Method 1: Use most recent term rate (Fall 22) • Method 2: Linear regression • Method 3: Random Forest

Algorithm Definition

- Dataframe

career_code	student_full_part_time	student_gender_ipeds	STUDENT_PROGRAM_TYPE	first_generation_code_adj	student_perm_county_rural_ind	fa_pell_offer_flag	Last term rate	Last 2 term rate	This term predicted rate
U	F	F	F2F	Y	Y	N	89%	91%	

.

.

.

Algorithm Definition

- Predict number of student returning in each group
- Sum the predicted count and calculate the retention rate
- Score is based on relative difference between predicted and actual retention rate
  - 0% difference -> 100% score
  - 20% difference ->0% score
- The idea: use three most recent years retention data as training set, and current year as test.

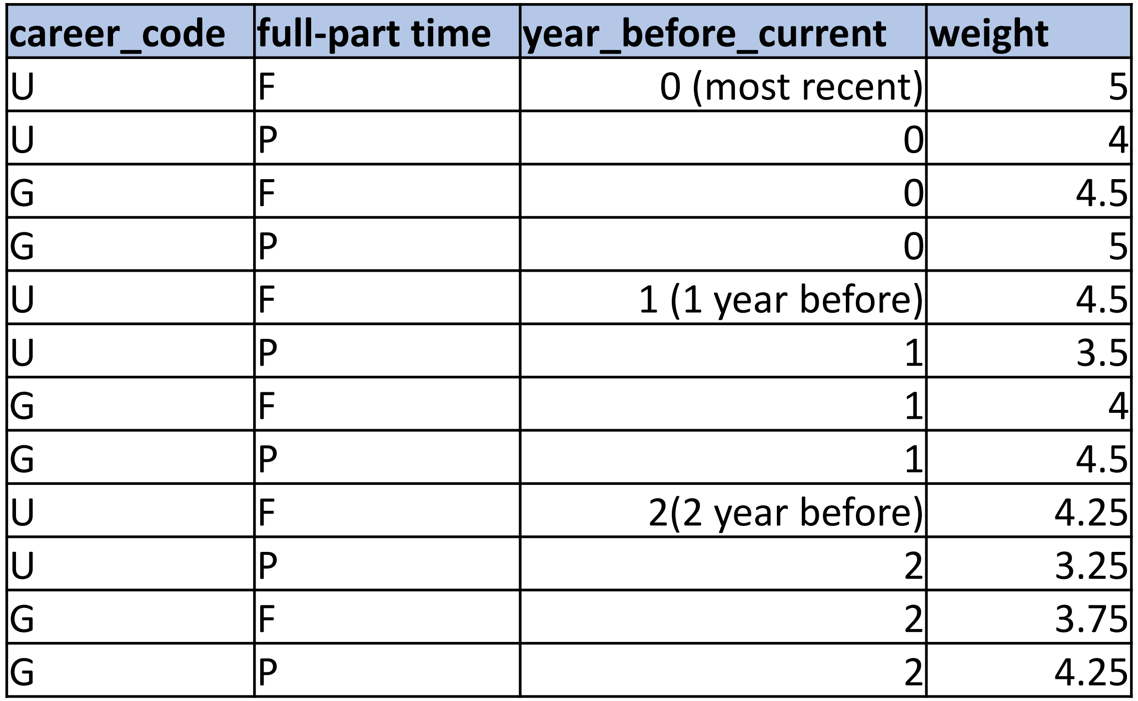
Algorithm Definition

- Define the criteria for model score
  - Four group of students: (undergraduate, graduate), (full time, part time)
  - Most recent 3 years

Parameters used in score criteria (career, full-part time) MUST be used in input ***(base parameters)***

Full-time UG and part-time GR have highest weighted

However, FT students have consistent retention rate, so more weighted can be put on PT students



career_code	full-part time	year_before_current	weight
U	F	0 (most recent)	5
U	P	0	4
G	F	0	4.5
G	P	0	5
U	F	1 (1 year before)	4.5
U	P	1	3.5
G	F	1	4
G	P	1	4.5
U	F	2(2 year before)	4.25
U	P	2	3.25
G	F	2	3.75
G	P	2	4.25

Algorithm Definition

- Example
  - Predict Fall 23 retention rate into Spring 24.
    - Input set A: career, full-part time, residency, Pell eligibility, class level
    - Input set B: career, full-part time, GPA category, military affiliated flag, academic standing

Term	Career	Full-part time	Actual retention rate	Predicted retention rate	
				Input A	Input B
Fall 2022	Undergraduate	Full time	90%	89%	90%
		Part time	75%	77%	76%
	Graduate	Full time	74%	73%	75%
		Part time	85%	86%	84%
Fall 2021	Undergraduate	Full time	87%	87%	87%
		Part time	78%	78%	78%
	Graduate	Full time	77%	77%	76%
		Part time	88%	87%	88%
Fall 2020	Undergraduate	Full time	89%	88%	88%
		Part time	80%	82%	80%
	Graduate	Full time	74%	75%	73%
		Part time	84%	82%	82%

Algorithm Definition

- Example
  - Determine model input to predict Fall 23 retention rate into Spring 24.
    - Calculate relative difference
    - Multiply by weight

Term	Career	Full-part time	Actual retention rate	Predicted retention rate		Rel. Difference (%)		Weight
				Input A	Input B	Input A	Input B	
Fall 2022	Undergraduate	Full time	90%	89%	90%	1.1%	0.0%	5
		Part time	75%	77%	76%	2.7%	1.3%	4
	Graduate	Full time	74%	73%	75%	1.4%	1.4%	4.5
		Part time	85%	86%	84%	1.2%	1.2%	5
Fall 2021	Undergraduate	Full time	87%	87%	87%	0.0%	0.0%	4.5
		Part time	78%	78%	78%	0.0%	0.0%	3.5
	Graduate	Full time	77%	77%	76%	0.0%	1.3%	4
		Part time	88%	87%	88%	1.1%	0.0%	4.5
Fall 2020	Undergraduate	Full time	89%	88%	88%	1.1%	1.1%	4.25
		Part time	80%	82%	80%	2.5%	0.0%	3.25
	Graduate	Full time	74%	75%	73%	1.4%	1.4%	3.75
		Part time	84%	82%	82%	2.4%	2.4%	4.25

Algorithm Definition

- Example
  - Compare score

Term	Career	Full-part time	Actual retention rate	Predicted retention rate		Rel. Difference (%)		Weight
				Input A	Input B	Input A	Input B	
Fall 2022	Undergraduate	Full time	90%	89%	90%	1.1%	0.0%	5
		Part time	75%	77%	76%	2.7%	1.3%	4
	Graduate	Full time	74%	73%	75%	1.4%	1.4%	4.5
		Part time	85%	86%	84%	1.2%	1.2%	5
Fall 2021	Undergraduate	Full time	87%	87%	87%	0.0%	0.0%	4.5
		Part time	78%	78%	78%	0.0%	0.0%	3.5
	Graduate	Full time	77%	77%	76%	0.0%	1.3%	4
		Part time	88%	87%	88%	1.1%	0.0%	4.5
Fall 2020	Undergraduate	Full time	89%	88%	88%	1.1%	1.1%	4.25
		Part time	80%	82%	80%	2.5%	0.0%	3.25
	Graduate	Full time	74%	75%	73%	1.4%	1.4%	3.75
		Part time	84%	82%	82%	2.4%	2.4%	4.25
					Score	93.9%	95.8%	

Algorithm Definition

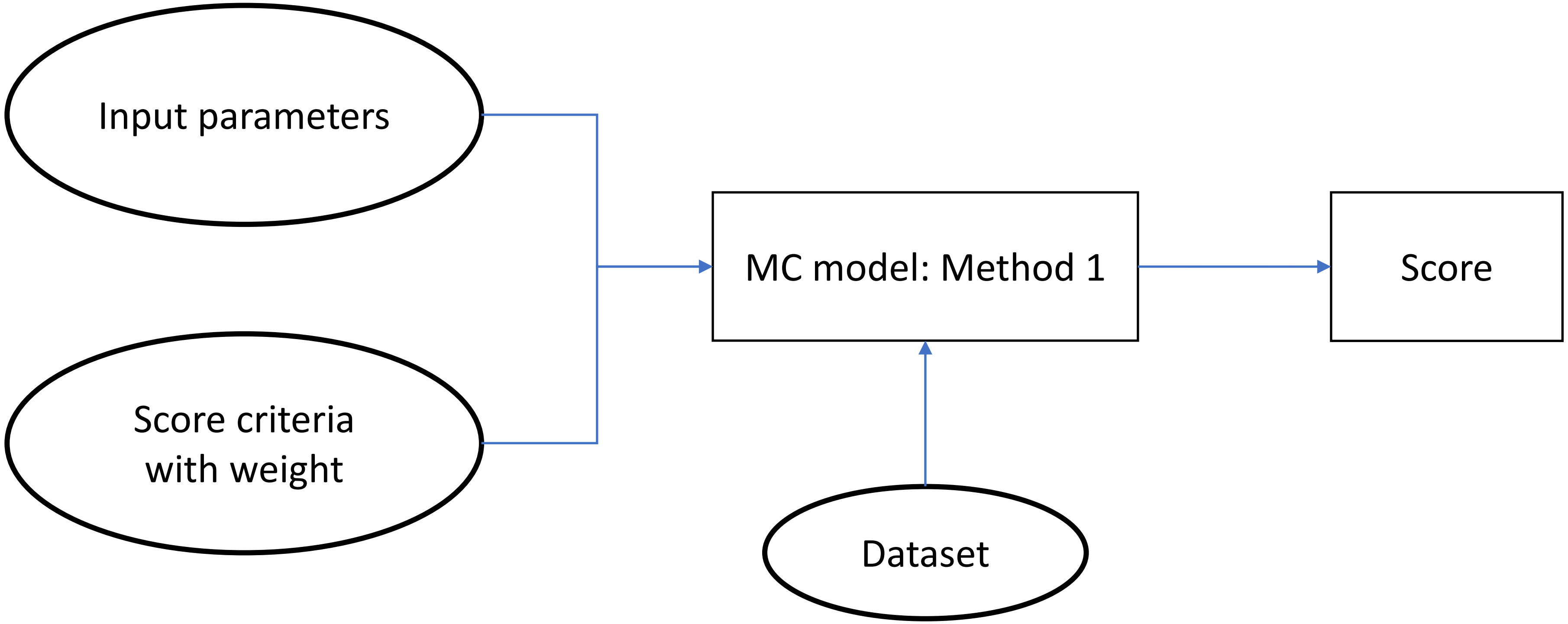
- Example
  - Compare score

Term	Career	Full-part time	Actual retention rate	Predicted retention rate		Rel. Difference (%)	
				Input A	Input B	Input A	Input B
Fall 2022	Undergraduate	Full time	90%	89%	90%	1.1%	0.0%
		Part time	75%	77%	76%	2.7%	1.3%
	Graduate	Full time	74%	73%	75%	1.4%	1.4%
		Part time	85%	86%	84%	1.2%	1.2%
Fall 2021	Undergraduate	Full time	87%	87%	87%	0.0%	0.0%
		Part time	78%	78%	78%	0.0%	0.0%
	Graduate	Full time	77%	77%	76%	0.0%	1.3%
		Part time	88%	87%	88%	1.1%	0.0%
Fall 2020	Undergraduate	Full time	89%	88%	88%	1.1%	1.1%
		Part time	80%	82%	80%	2.5%	0.0%
	Graduate	Full time	74%	75%	73%	1.4%	1.4%
		Part time	84%	82%	82%	2.4%	2.4%
					Score	93.9%	95.8%

Use input set B for Fall 23 prediction

Algorithm Workflow

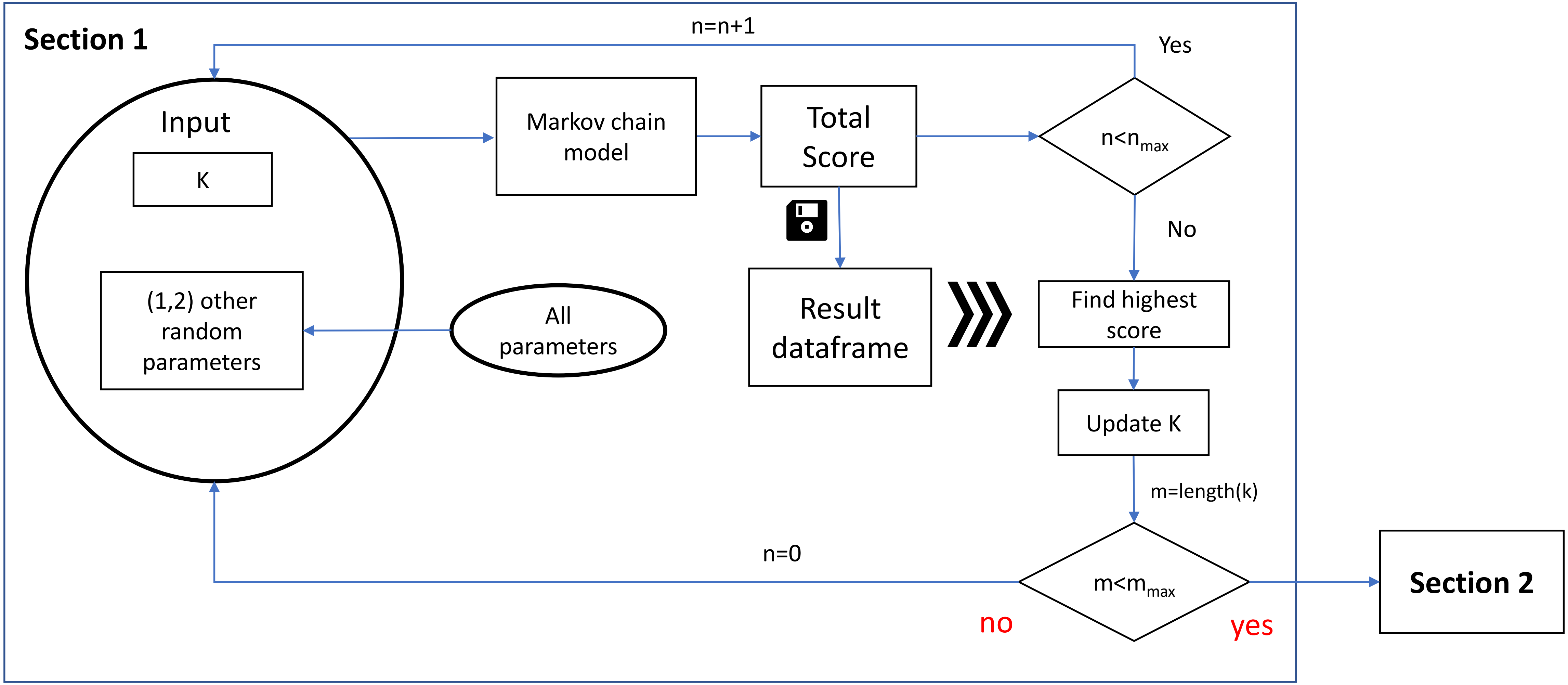
- Markov chain model used in the algorithm



Algorithm Workflow

- Define constants
  - K: list of parameters, starting with (career_code, full_part_time)
  - n_{max}: maximum number of iterations per cycle (default 200)
  - m_{max}: maximum number of parameters used for prediction *length*(*K*) (8-12)

Algorithm Workflow



Section 2

- Result dataframe format

Career_code	Full-part_time	Class_level	residency	Pell_eligibility	GPA_cat	Enrll_stt	gender	….	Score
1	1	1	0	1	0	0	1		91
1	1	0	1	0	1	1	0		92
1	1	1	0	0	1	0	1		89

- Perform logistics regression

Section 2

- Each parameter should have coefficient and p-value

Parameters	Coefficient	P-value
residency	0.2	<1e-9
gender	0.5	0.5
race_ipeds	-1	0.0001
class_level	1	<1e-9
enrollment_status	-0.6	1
…		

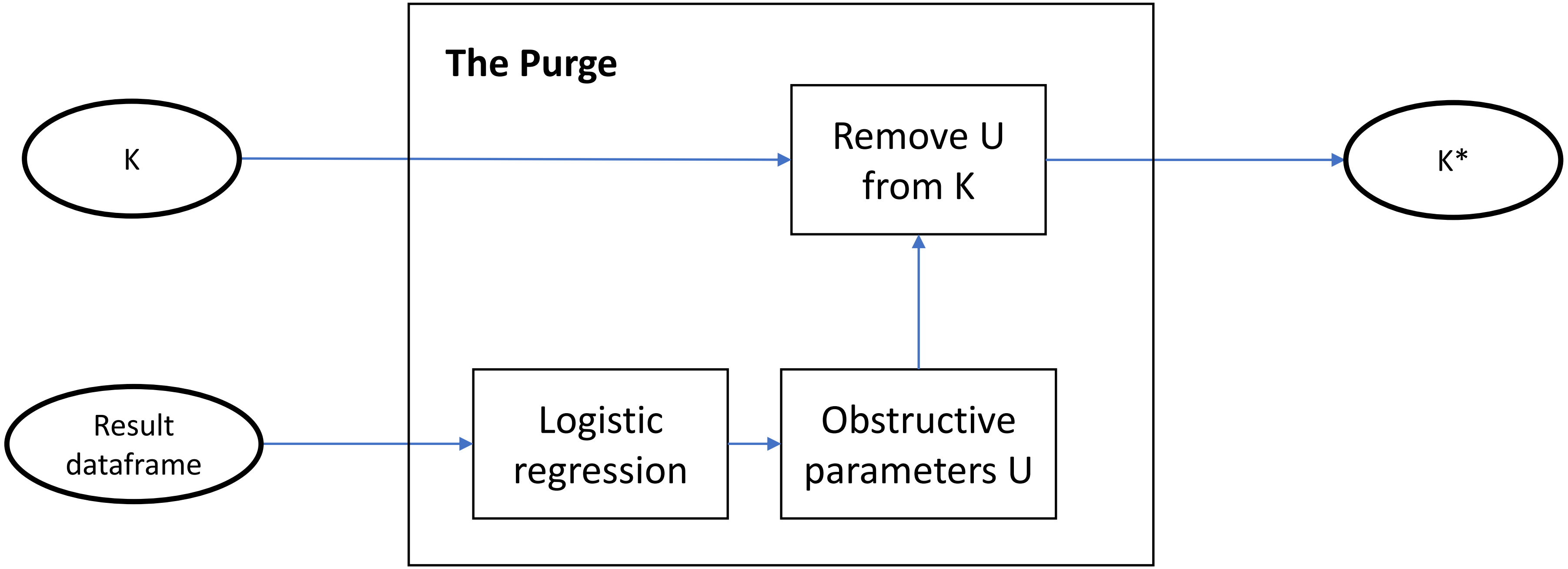
Section 2

- Select parameters with negative coefficient and statistically significant

Parameters	Coefficient	P-value	Statistically significant?	Negative coefficient
residency	0.2	<1e-5	Y	N
gender	0.5	0.5	N	N
race_ipeds	-1	0.0001	Y	Y
class_level	1	<1e-5	Y	N
enrollment_status	-0.6	1	N	Y
…				

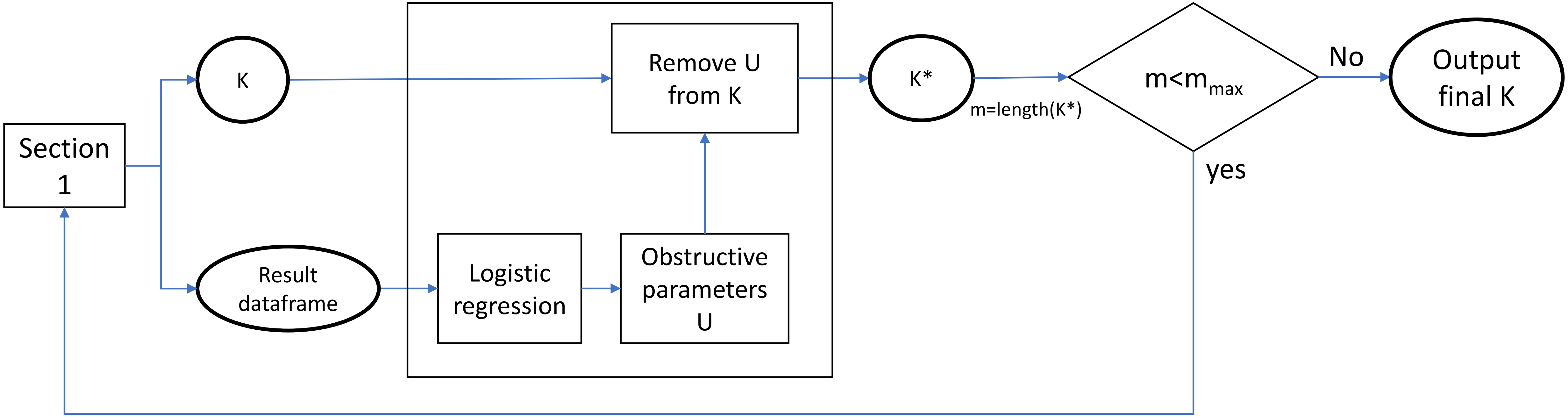
Model Workflow

- Combine list of “undesirable parameters”
- Remove these parameters from list of parameters K (if any)



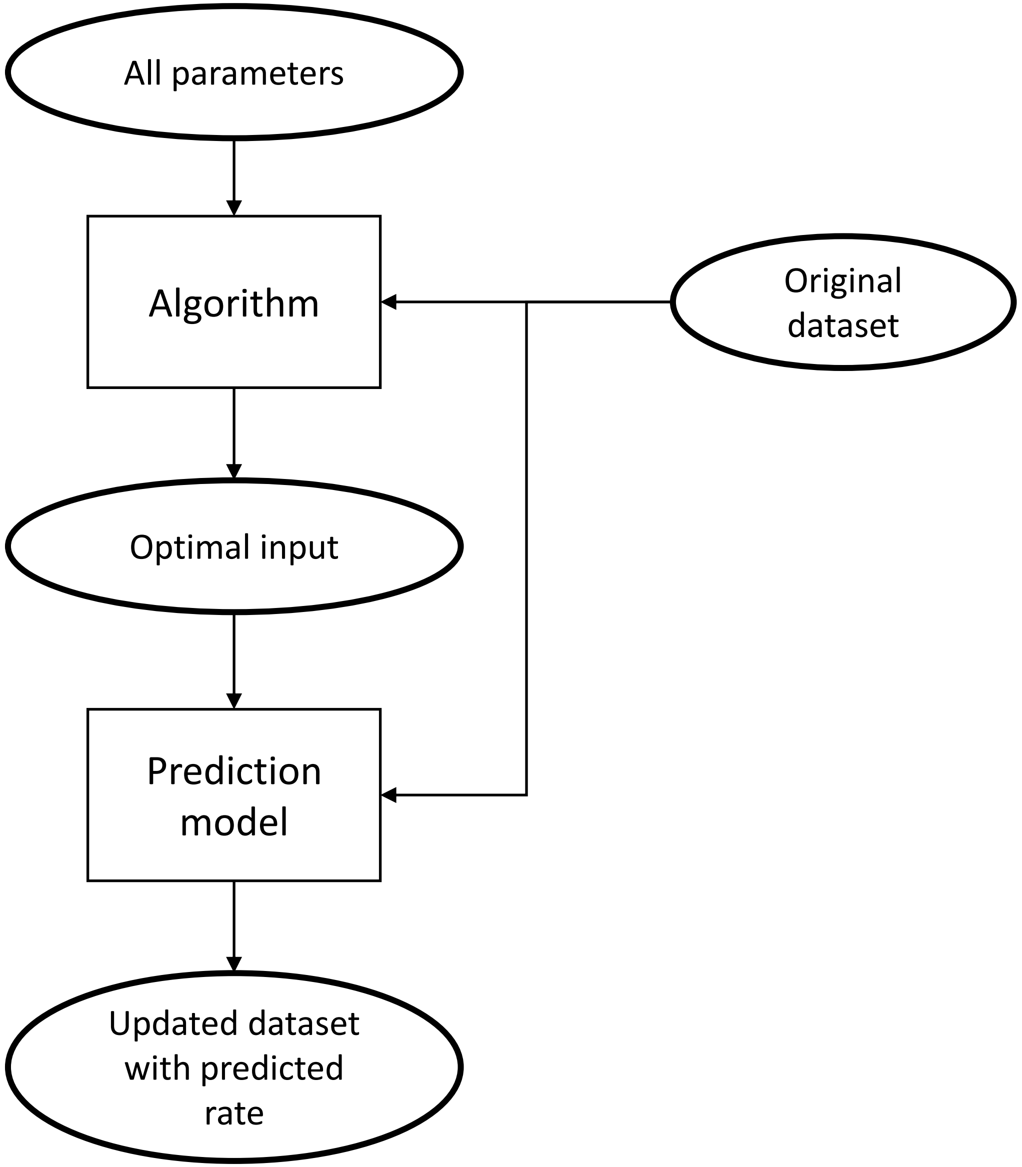
Model Workflow

- Combine list of “undesirable parameters”
- Remove these parameters from list of parameters K (if any)



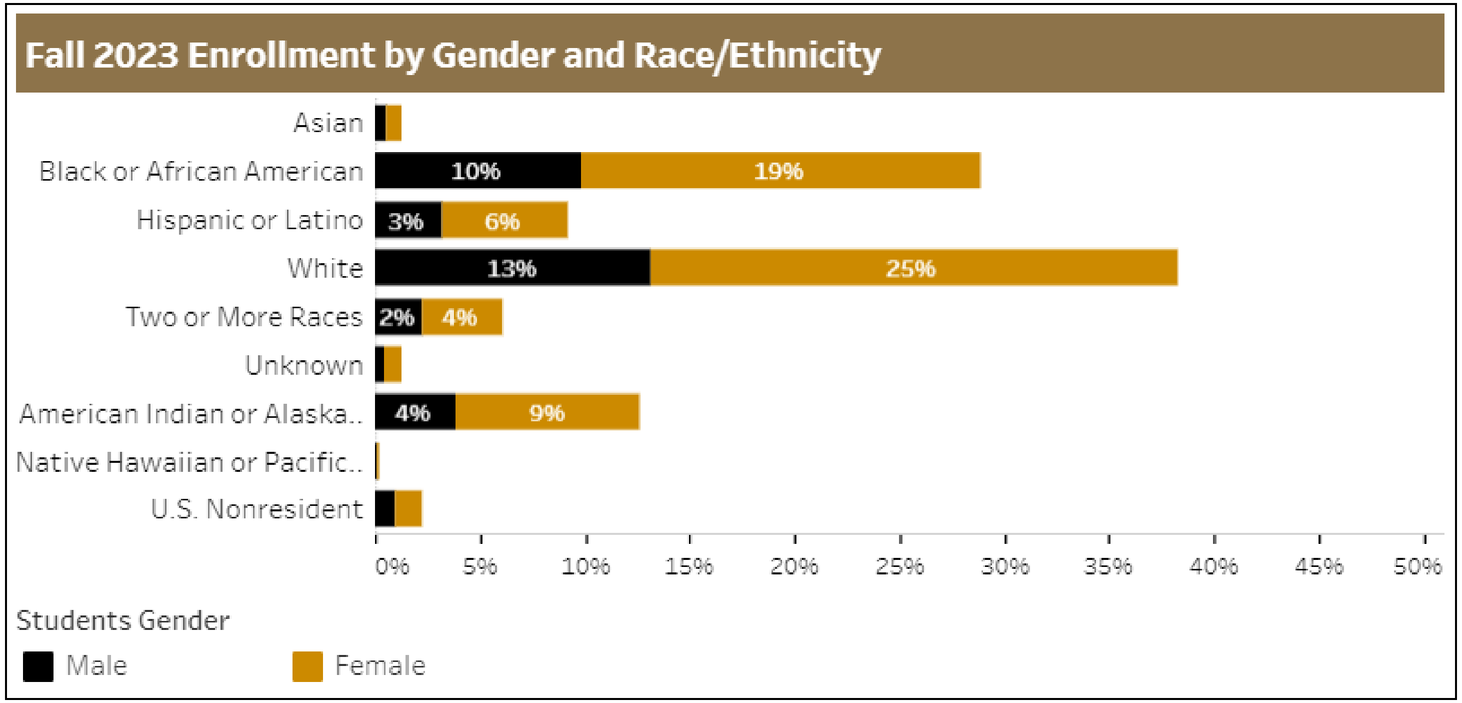
Model Workflow

- Final workflow
- The dataset can be loaded into Tableau for visualization and in-depth analysis

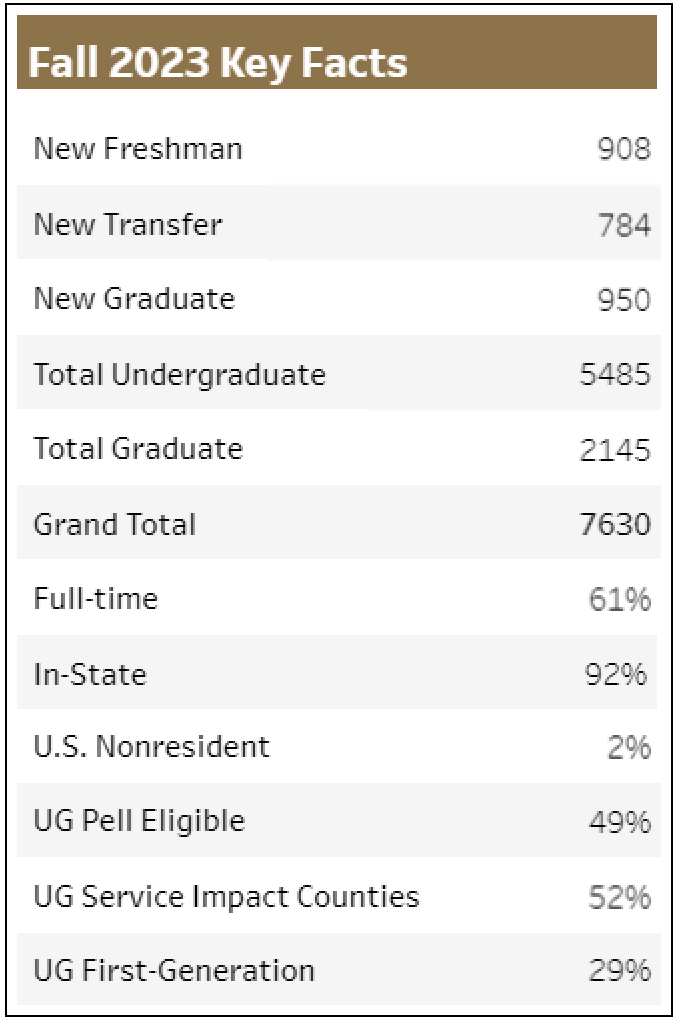


Overview of Data

- Overview of UNCP



Fall 2023 Enrollment by Gender and Race/Ethnicity

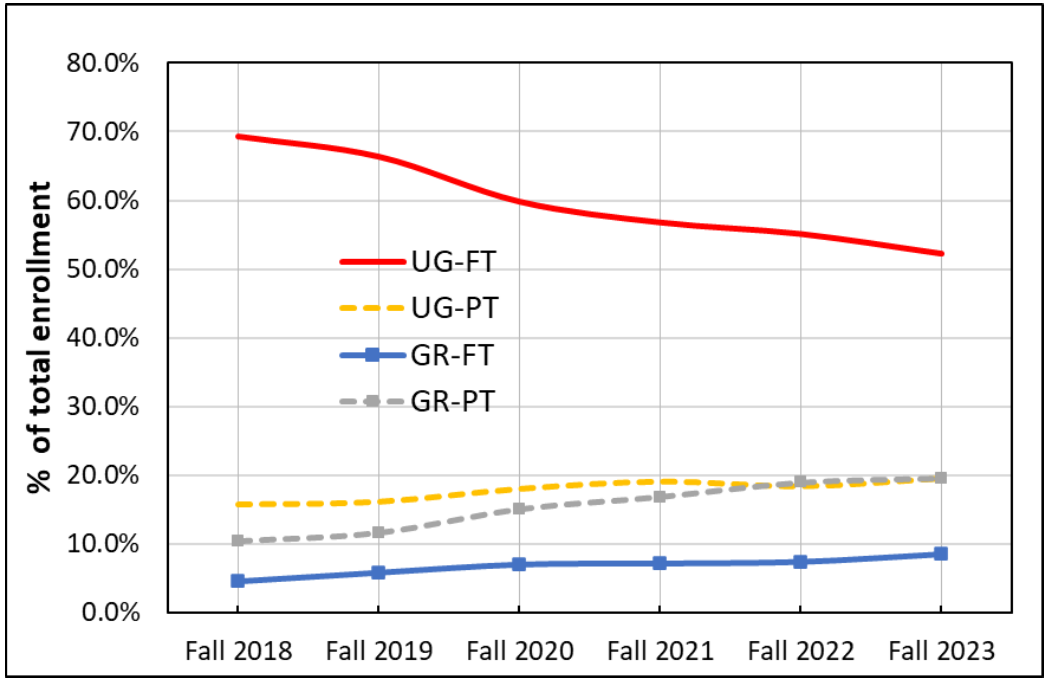


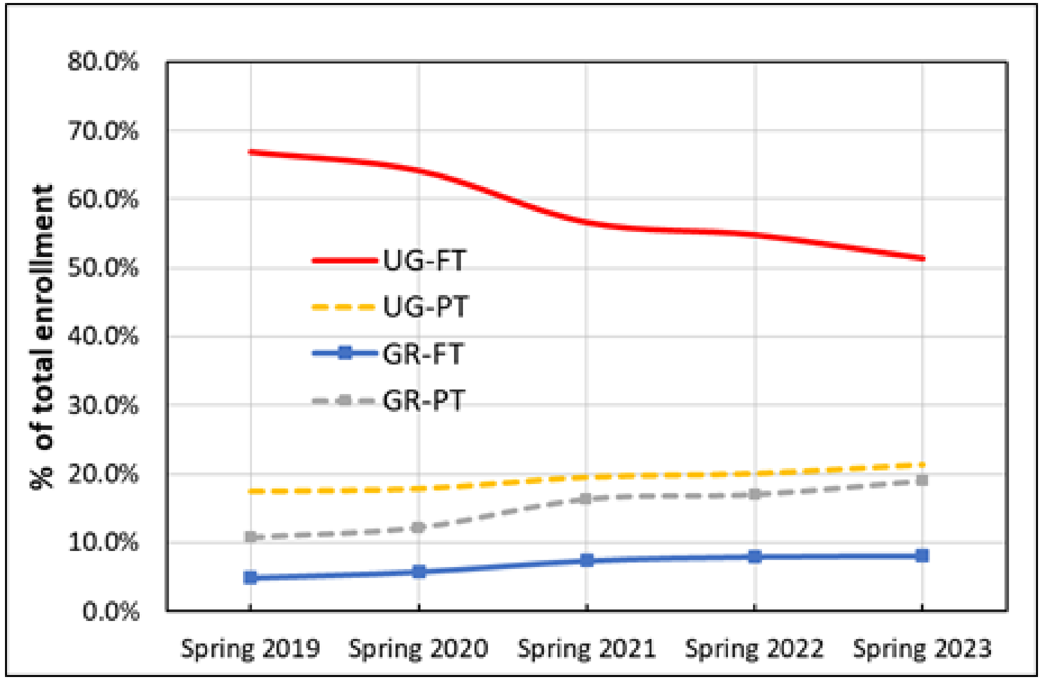
Fall 2023 Key Facts

New Freshman	908
New Transfer	784
New Graduate	950
Total Undergraduate	5485
Total Graduate	2145
Grand Total	7630
Full-time	61%
In-State	92%
U.S. Nonresident	2%
UG Pell Eligible	49%
UG Service Impact Counties	52%
UG First-Generation	29%

Overview of Data

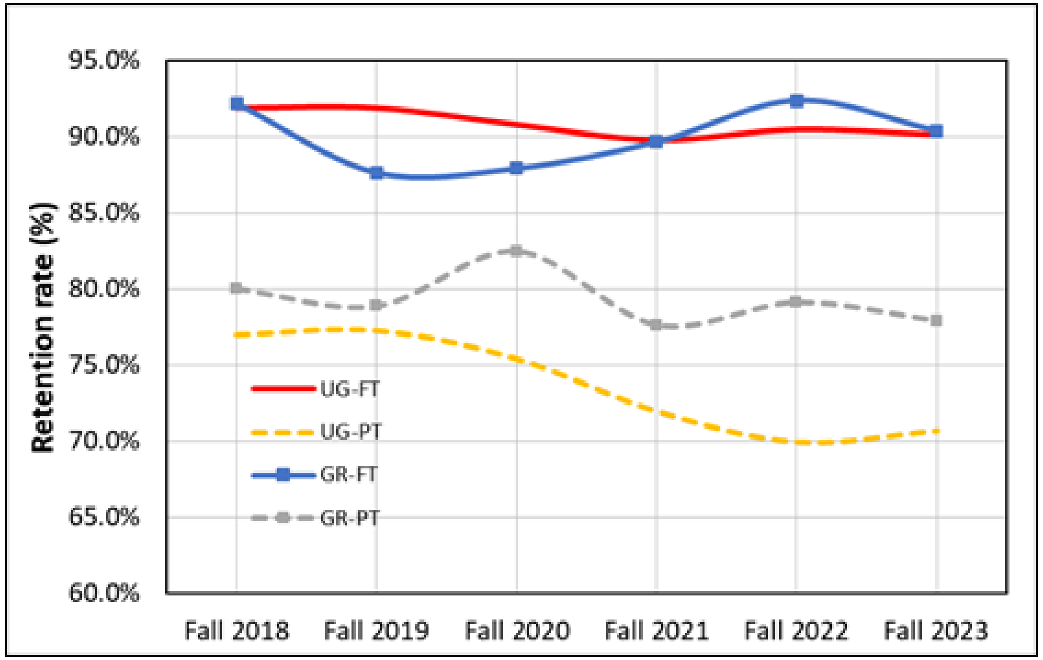
- Enrollment by groups

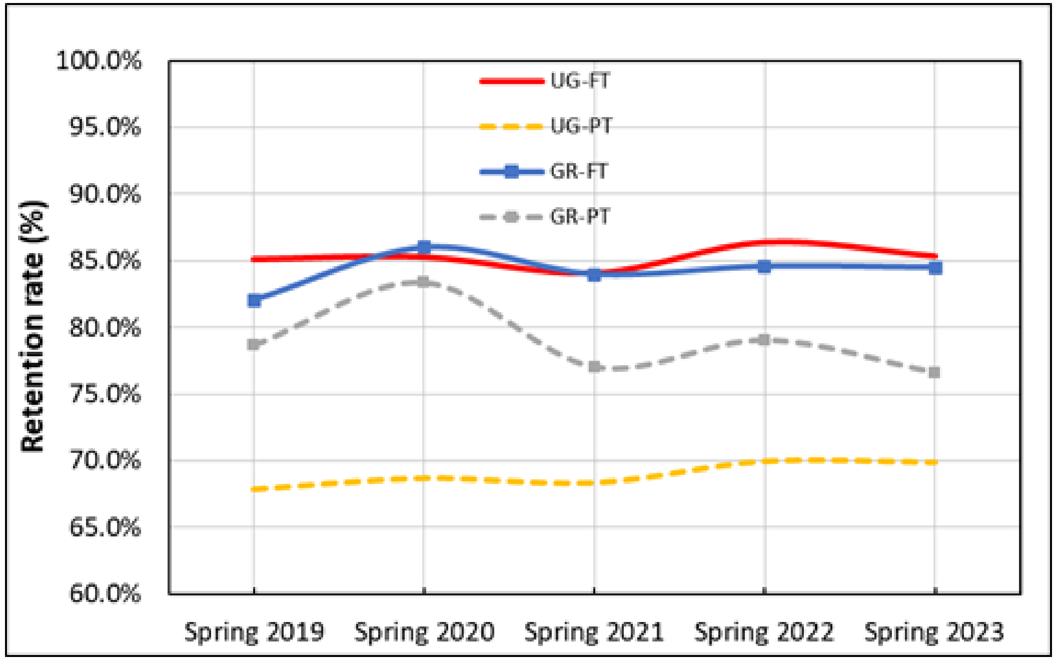




Overview of Data

- Retention by groups





Results

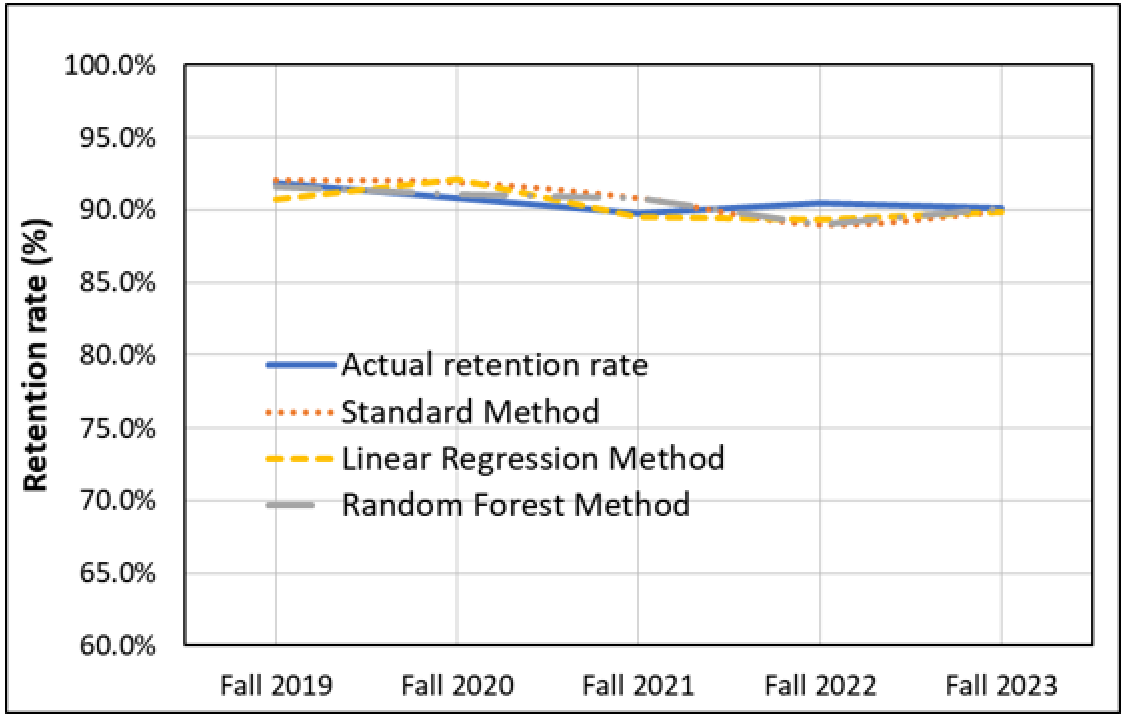
- Example of best parameters for Fall 23 undergraduate

{"career_code", "student_full_part_time", "unmet_need_flag", "enrollment_status_short", "full_term_flag", "prev_term_dfwi_flag", "have_lecture_flag", "athlete_flag", "normal_astd_bot_flag", "priority_fafsa_time_met"}

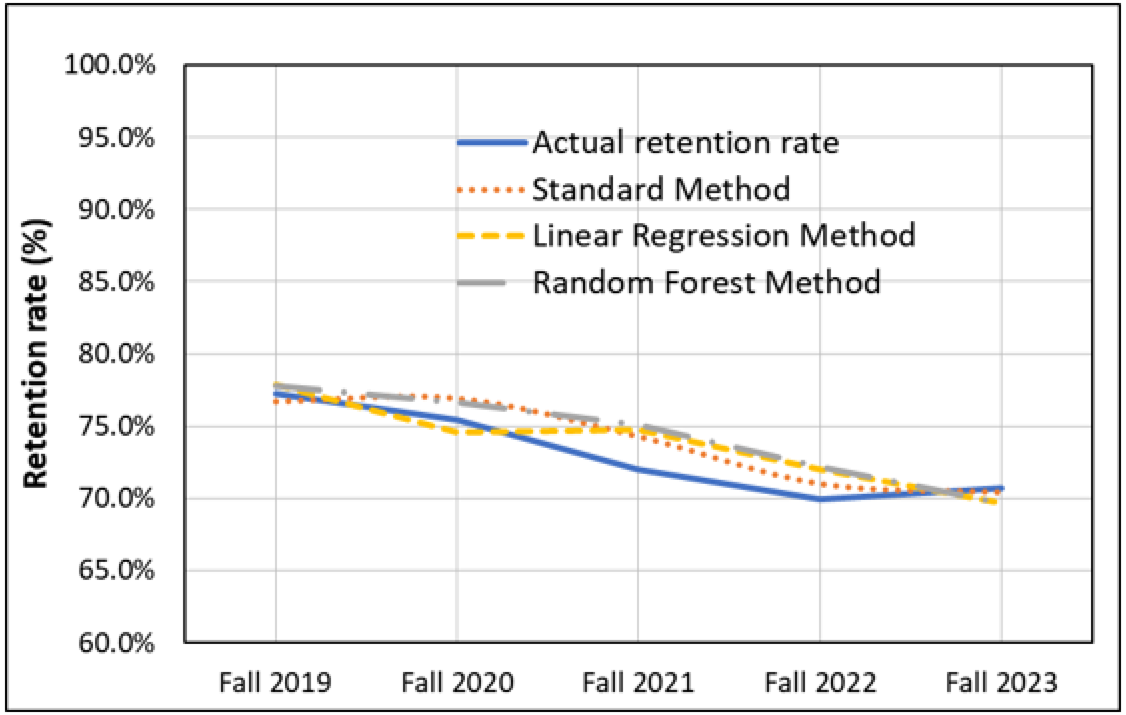
Term period	Term	Training Score	
		Undergraduate	Graduate
Fall	Fall 2019	90.3	95.1
	Fall 2020	94.5	98.5
	Fall 2021	98.9	95.8
	Fall 2022	96.7	86.8
	Fall 2023	98.3	91.6
Spring	Spring 2020	98.1	92.7
	Spring 2021	99.2	89.5
	Spring 2022	99.7	87.2
	Spring 2023	96.0	92.3

Results

- Fall-to-spring retention
- Undergraduate



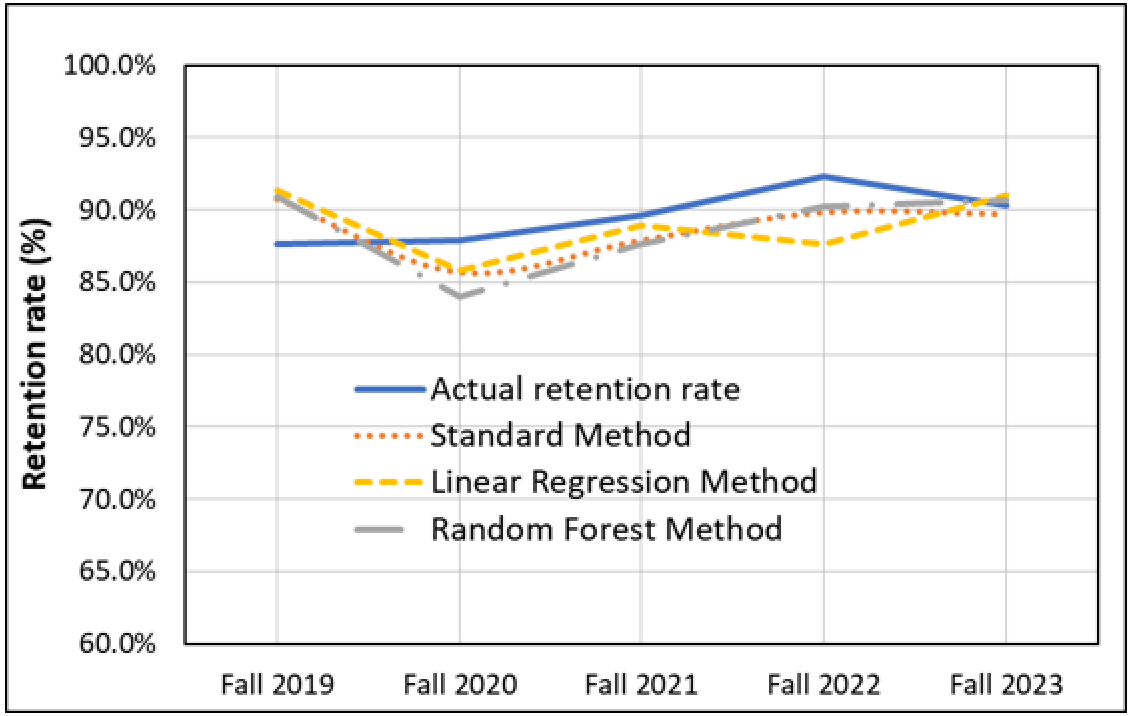
a. UG-FT



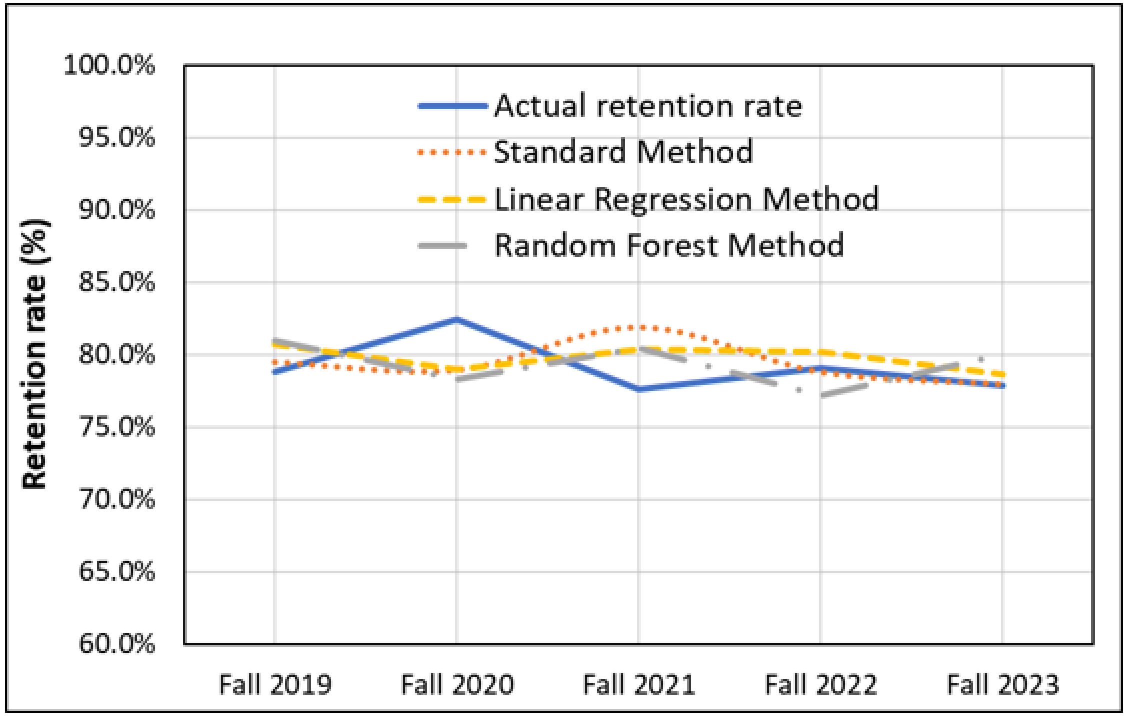
b. UG-PT

Results

- Fall-to-spring retention
- Graduate



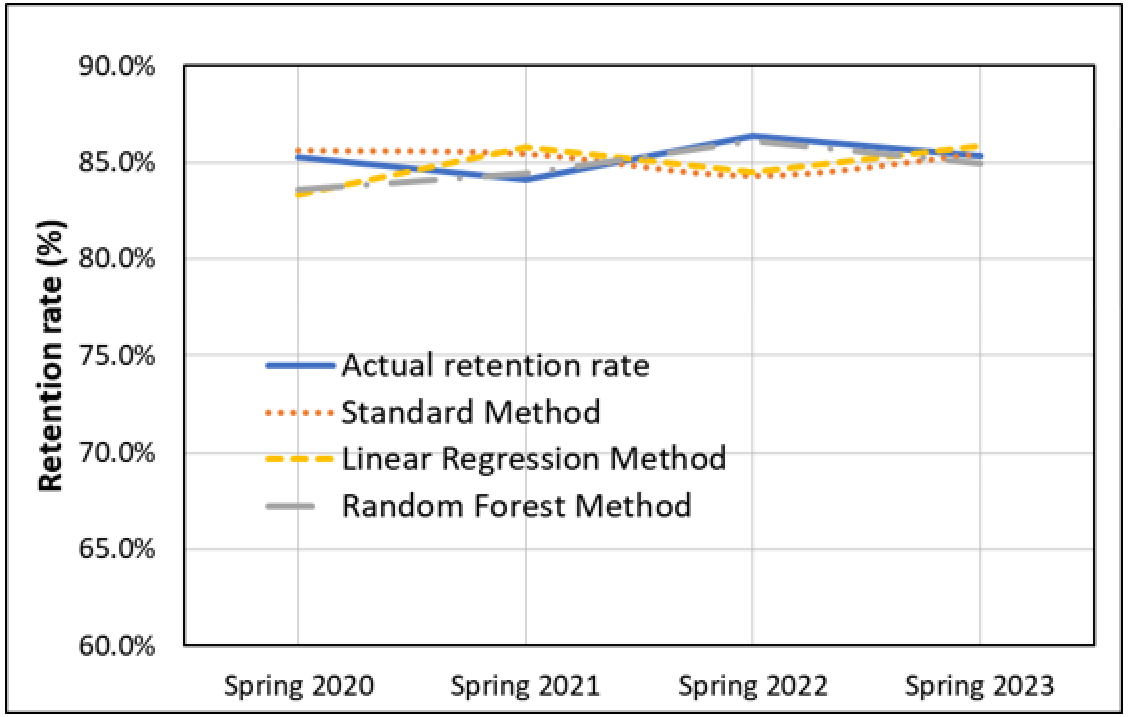
c. GR-FT



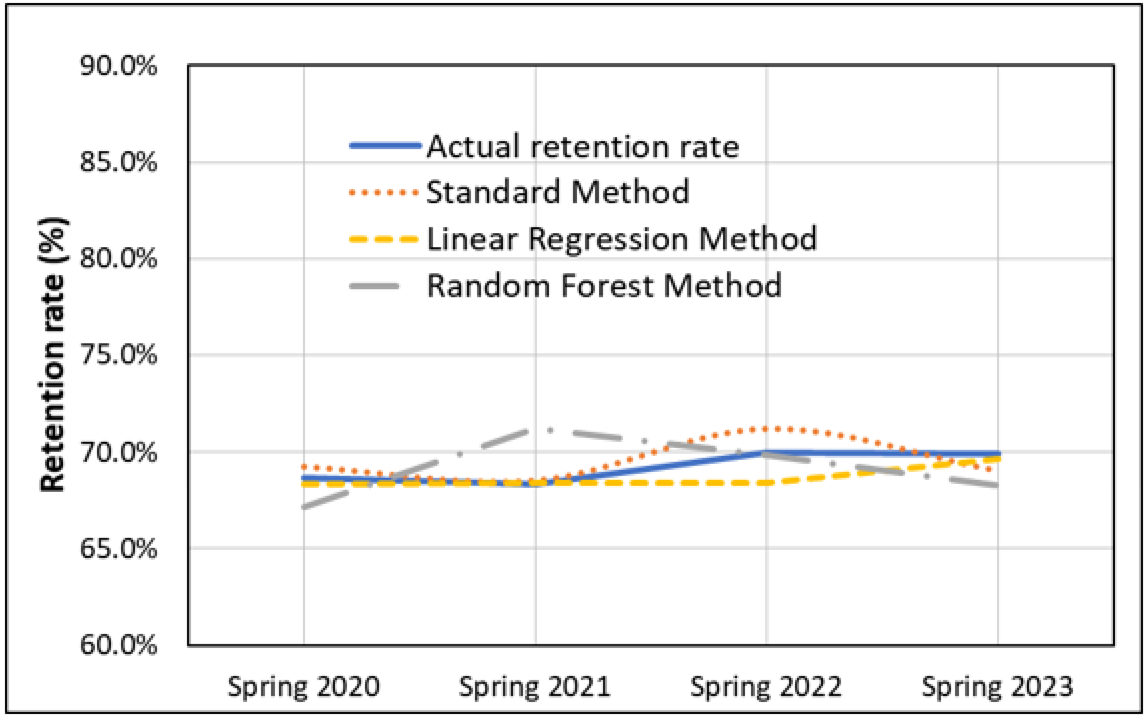
d. GR-PT

Results

- Spring-to-fall retention
- Undergraduate



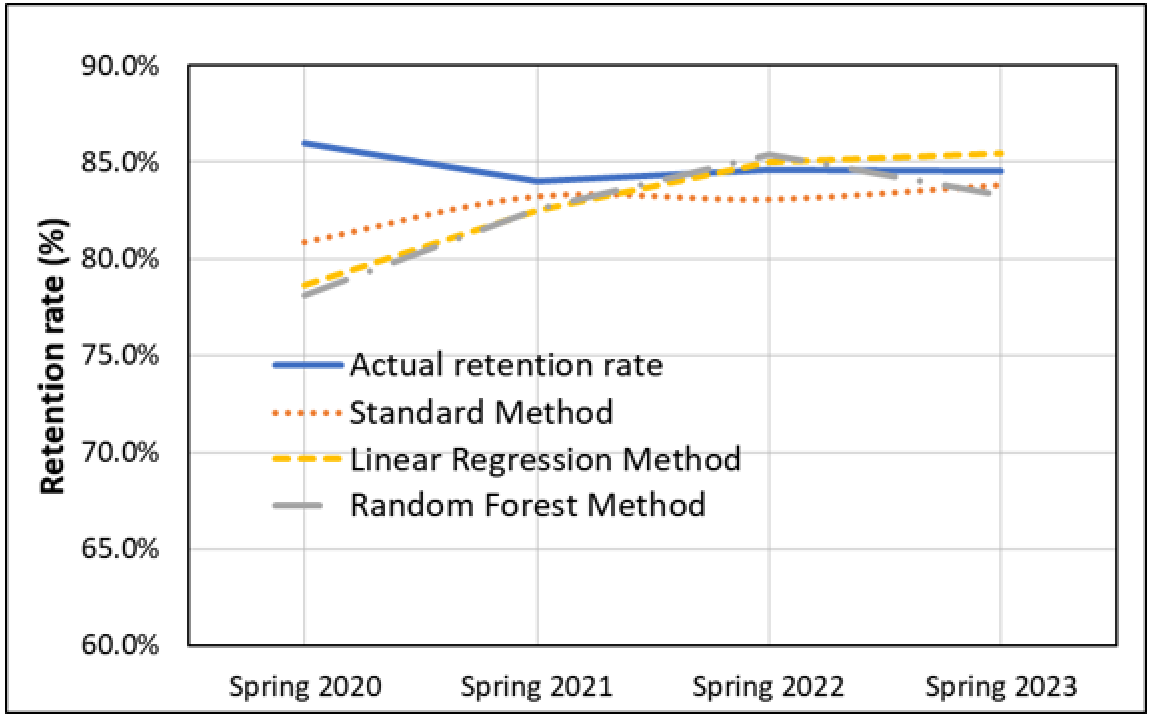
a. UG-FT



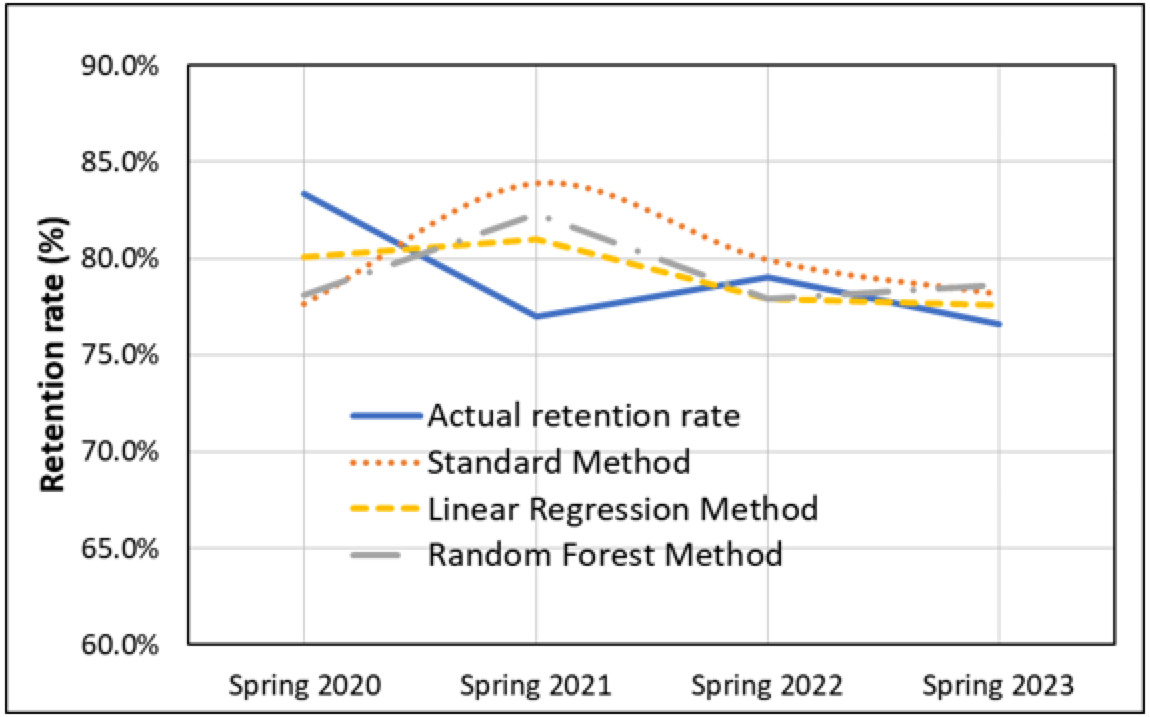
b. UG-PT

Results

- Spring-to-fall retention
- Graduate



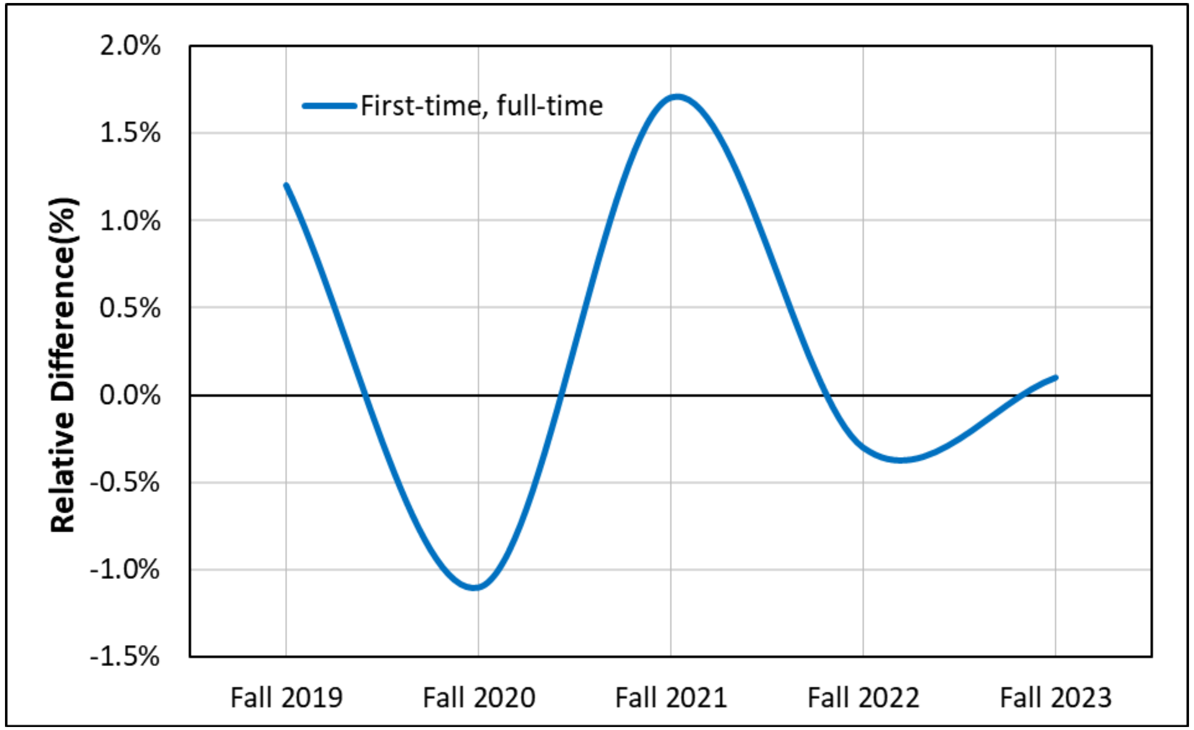
c. GR-FT



d. GR-PT

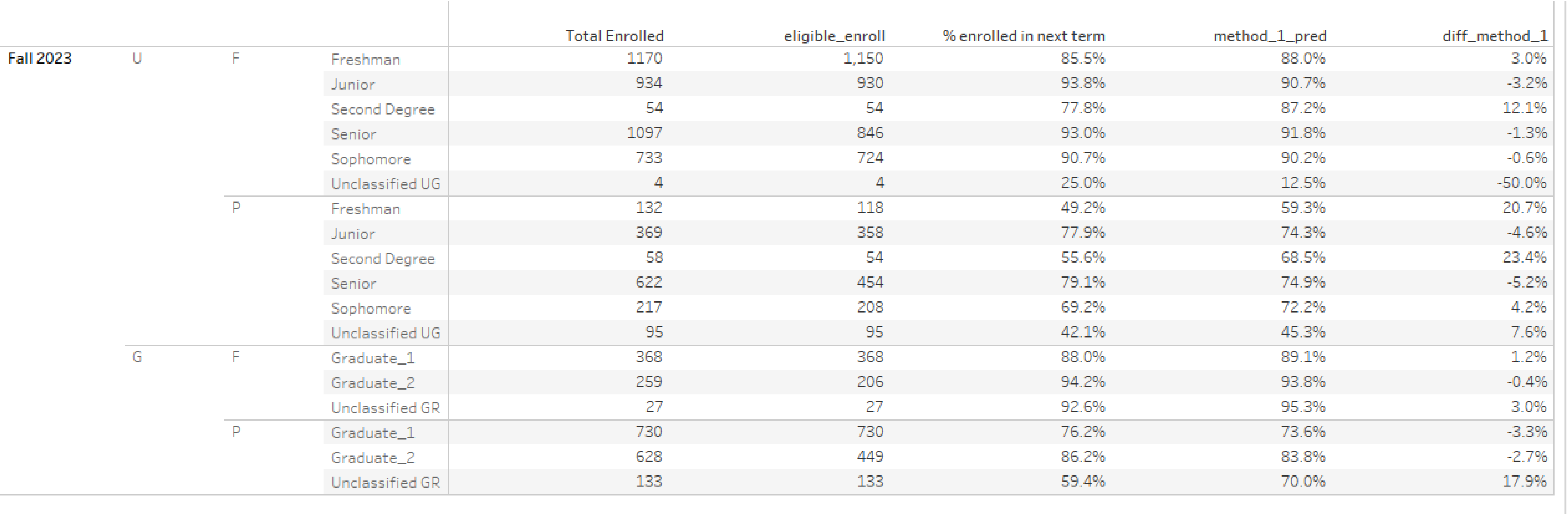
Results

- Fall-to-spring retention of FTFT students



Results

- Example of in-depth analysis using Tableau



				Total Enrolled	eligible_enroll	% enrolled in next term	method_1_pred	diff_method_1
Fall 2023	U	F	Freshman	1170	1,150	85.5%	88.0%	3.0%
			Junior	934	930	93.8%	90.7%	-3.2%
			Second Degree	54	54	77.8%	87.2%	12.1%
			Senior	1097	846	93.0%	91.8%	-1.3%
			Sophomore	733	724	90.7%	90.2%	-0.6%
			Unclassified UG	4	4	25.0%	12.5%	-50.0%
		P	Freshman	132	118	49.2%	59.3%	20.7%
			Junior	369	358	77.9%	74.3%	-4.6%
			Second Degree	58	54	55.6%	68.5%	23.4%
			Senior	622	454	79.1%	74.9%	-5.2%
			Sophomore	217	208	69.2%	72.2%	4.2%
			Unclassified UG	95	95	42.1%	45.3%	7.6%
	G	F	Graduate_1	368	368	88.0%	89.1%	1.2%
			Graduate_2	259	206	94.2%	93.8%	-0.4%
			Unclassified GR	27	27	92.6%	95.3%	3.0%
		P	Graduate_1	730	730	76.2%	73.6%	-3.3%
			Graduate_2	628	449	86.2%	83.8%	-2.7%
			Unclassified GR	133	133	59.4%	70.0%	17.9%

Conclusion

- The proposed algorithm-model can provide more accurate prediction of term-to-term retention
- Useful in enrollment prediction and detection of abrupt change in behavior
- Easy to interpret and visualize
- More advanced predictive algorithm can be integrated to the model
- However, it cannot take into account unforeseeable factors (COVID)
- Part-time students behavior is sporadic, may need more tailored approach
- Future work aims to include FTE prediction and class level

References

- Eric Yang, Dantong Yang. *Markov Model with Survey Data to Project Enrollment during the Pandemic*. AIR Forum Virtual 2021.
- Samantha Bradley. *Predicting Student Enrollment Using Markov Chain Modeling in SAS*. SAIR 2018 Pre-Conference Workshop.
- Rick Sears. Forecasting Seat Demand. SAIR 2022.
- Brezavšček, Alenka, Bach, Mirjana Pejić and Baggia, Alenka. "Markov Analysis of Students' Performance and Academic Progress in Higher Education" Organizacija, vol.50, no.2, 2017, pp.83-95. https://doi.org/10.1515/orga-2017-0006
- Fatima, Taif, Marhane Khaoula, and Namir Abdelwahed. "Predicting Number of Student Enrollment Using a Discrete-Time Markov Chain Model." International Conference on Advanced Intelligent Systems for Sustainable Development. Cham: Springer International Publishing, 2020.
- Keener, Thomas. Analysis of College Graduation Rates Using Markhov Chains. Diss. Appalachian State University, 2022.
- Gandy, Rex; Crosby, Lynne; Luna, Andrew; Kasper, Daniel; Kendrick, Sherry. " Enrollment Projection Using Markov Chains: Detecting Leaky Pipes and the Bulge in the Boa". The AIR Professional File, Fall 2019. Article 147.
- Jie Dai, Li An. *1.21 - Time Geography*. Comprehensive Geographic Information Systems, Elsevier, 2018, Pages 303-312. ISBN 9780128047934. https://doi.org/10.1016/B978-0-12-409548-9.09625-1.

Appendix

Parameters	Description	Source
career_code	Indicates whether the student is undergraduate or graduate	SDM
student_full_part_time	Indicates whether the student is full time or part time	SDM
student_gender_ipeds	Male/female/Other	SDM
STUDENT_PROGRAM_TYPE	Indicates whether the student enrolls in Face-to-face or Online program	Banner
first_generation_code_adj	Indicates whether the student is first generation student	SDM
student_perm_county_rural_ind	Indicates whether the student's permanent county is considered rural by state government	SDM
fa_pell_offer_flag	Indicates whether the student received Pell offer status for the semester	SDM
enrollment_status_short	Classification of enrollment status: 1. New Freshmen 2. New Transfer 3. Continuing Undergraduate 4. Non-degree Students 5. New Graduate Students 6. Continuing Graduate	SDM
fte_cat	Student's FTE value from 0.25 to 1 based on course load	SDM
residency	Indicates whether the student is In-state/out-of-state for tuition purpose	SDM
und_races_flag	Indicates whether the student's race and ethnicity is considered underrepresented (Black, Hispanic, Native American..)	SDM
normal_astd_bot_flag	Indicates whether the student is in normal academic standing at start of the semester (no warning, probation)	Banner
cgpa_cat	Cumulative GPA category 1. Not available (for new/non-degree seeking students) 2. Below 2.0 3. From 2.0 to 3.0 4. Above 3.0	SDM
military_affiliated_flag	Indicates whether the student is military affiliated (including dependent)	Banner
hold_flag	Indicates whether the student has registration holds	Banner
prev_term_enrl_flag	Y/N for students enrolling/not enrolling in previous Fall/Spring NA for new students	SDM
adult_learner_flag	Indicates whether the student is over the age of 24	SDM

Thank you for attending the 2024 NCAIR Annual Conference!

There's a QR code in your program for a conference evaluation form. You'll also get an e-mail following the conference with a link to the form, which will be available until 4/30.

At your earliest convenience, please take this opportunity to let the planning committee know your thoughts about this year's conference and where you would like to meet next year.