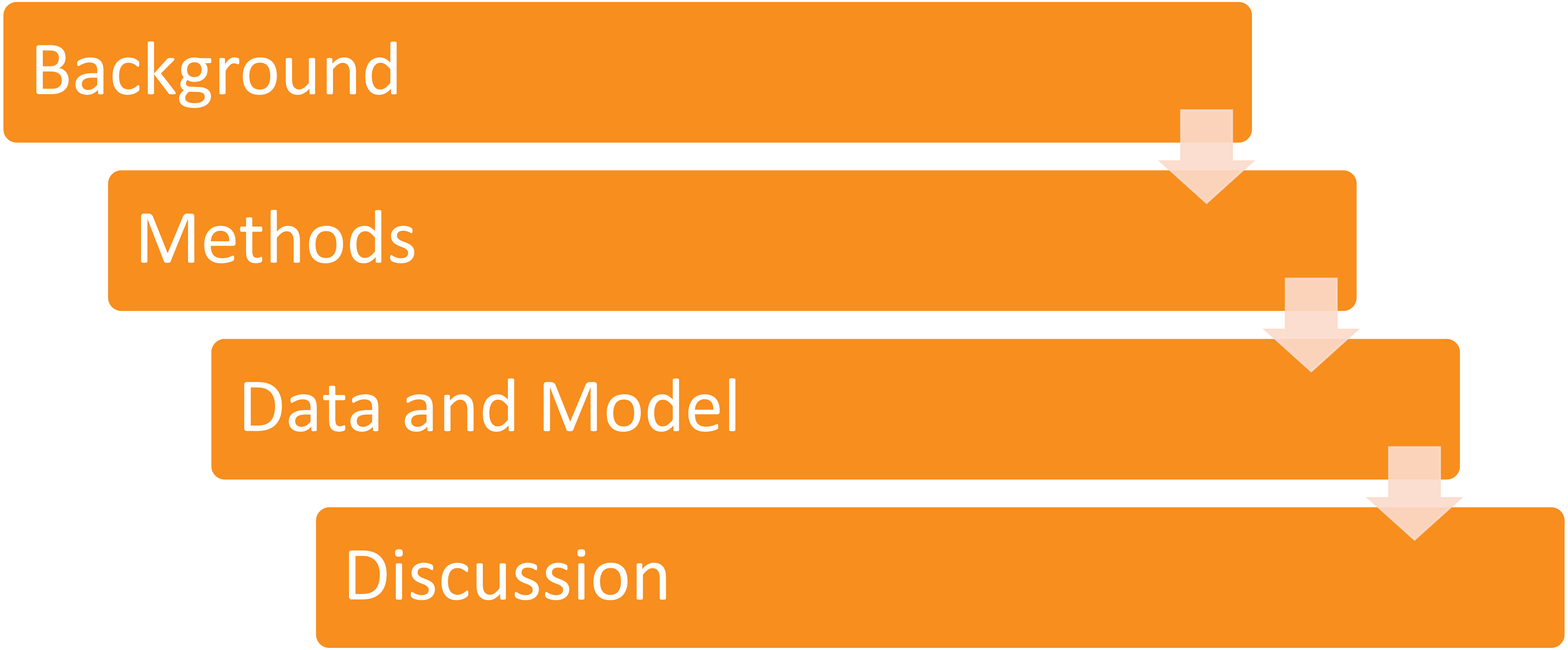
Predicting First-Time Freshman Retention with Pre-Attendance Data

Kristina Beltran, Brandon Cooper, & Tristan Young

Outline



Background

Methods

Data and Model

Discussion

Background

First-Time Freshman Retention

Sam Houston State University (SHSU) aims to boost its First-Time Freshman (FTF) retention from the current rate of ~75%.

Early Intervention

Next-generation advising seeks to "proactively identify and connect with students who [are] struggling" (Faugh 2023:2).

Cross-functional Collaboration

Academic Affairs, Data Analytics & Decision Support, and others join forces to direct their resources on the problem.

Variables Considered

Feature	Definitions
one_yr_ret_ind	Indicator (Y/N) for retained
ethnicity	Multilevel factor following IPEDS definitions
gender	Multilevel factor following THECB definitions
first_gen	Neither parent nor guardian has earned a bachelor's degree in the US
college	College of major field of study
feeder_ind	Indicator for high schools that send ≥10 students over any three years in a five-year period
high_school_gpa	High school GPA from application
sat_concordance	SAT concordance score
received_pell	Indicator (Y/N) for receiving any amount of a Pell grant in their fall FTF semester.
federal_aid	Numerical value of federal aid awarded, no loans or work study funds were included
state_aid	Numerical value of state aid awarded, no loans or work study funds were included
institutional_aid	Numerical value of institutional aid awarded, no loans or work study funds were included
private_aid	Numerical value of private aid awarded, no funds from loans were included
loan_aid	Numerical value of loan aid awarded, no grants or scholarships funds were included
work_study_aid	Numerical value of federal or state work study aid awarded
no_aid	Indicator (Y/N) for received no aid

Variables Considered, ctd.

Feature	Definitions
agi	Numerical value of adjusted gross income
fall_credits_attempted	Numerical value for total credit hours attempted
tsi_math	Indicator (Y/N) for TSI math complete by census day of the fall FTF term
tsi_writ	Indicator (Y/N) for TSI writing complete by census day of the fall FTF term
tsi_read	Indicator (Y/N) for TSI reading complete by census day of the fall FTF term
app_day_from_cutoff	Numerical value derived by the number of days from date of submitted application until the date of application close
athlete_ind	Indicator (Y/N) for student athlete
emp_before_term_ind	Employed on campus before term start (Y/N)
prereg_workload_answer	Multilevel factor for hours and location (on-/off-campus) of employment
prereg_residence_answer	Multilevel factor for residence (in Walker County; on- or off-campus; and with/without family)

Methods

Classification

Predicting FTF retention is a classification problem.

Utilization of Random Forest Model

Random Forest modeling is one approach used for determining variable importance and for prediction.

Inclusion of Pre-University Factors

To assist with early intervention, factors available before the first class day are considered in this analysis.

Data, Handling, & Analysis

Data

17,784 FTF included (AYs 2017-2023).
26 variables analyzed.

Exploration and Handling

NAs retained through indicators, allowing us to utilize missing data as information in itself. Sample balanced.

Analysis

Used Python SciKit-Learn. Variables removed through backward selection, multicollinearity, low category counts, and 50/50 split of outcomes.

Student Population

Cohort Year	Total FTF Students	% Female	% Minority	% First-Gen
Fall 2017	2,854	63	48	50
Fall 2018	2,870	63	50	52
Fall 2019	2,916	63	51	49
Fall 2020	2,842	64	52	42
Fall 2021	2,908	65	50	26
Fall 2022	3,394	63	50	46
Fall 2023	3,600	64	54	49

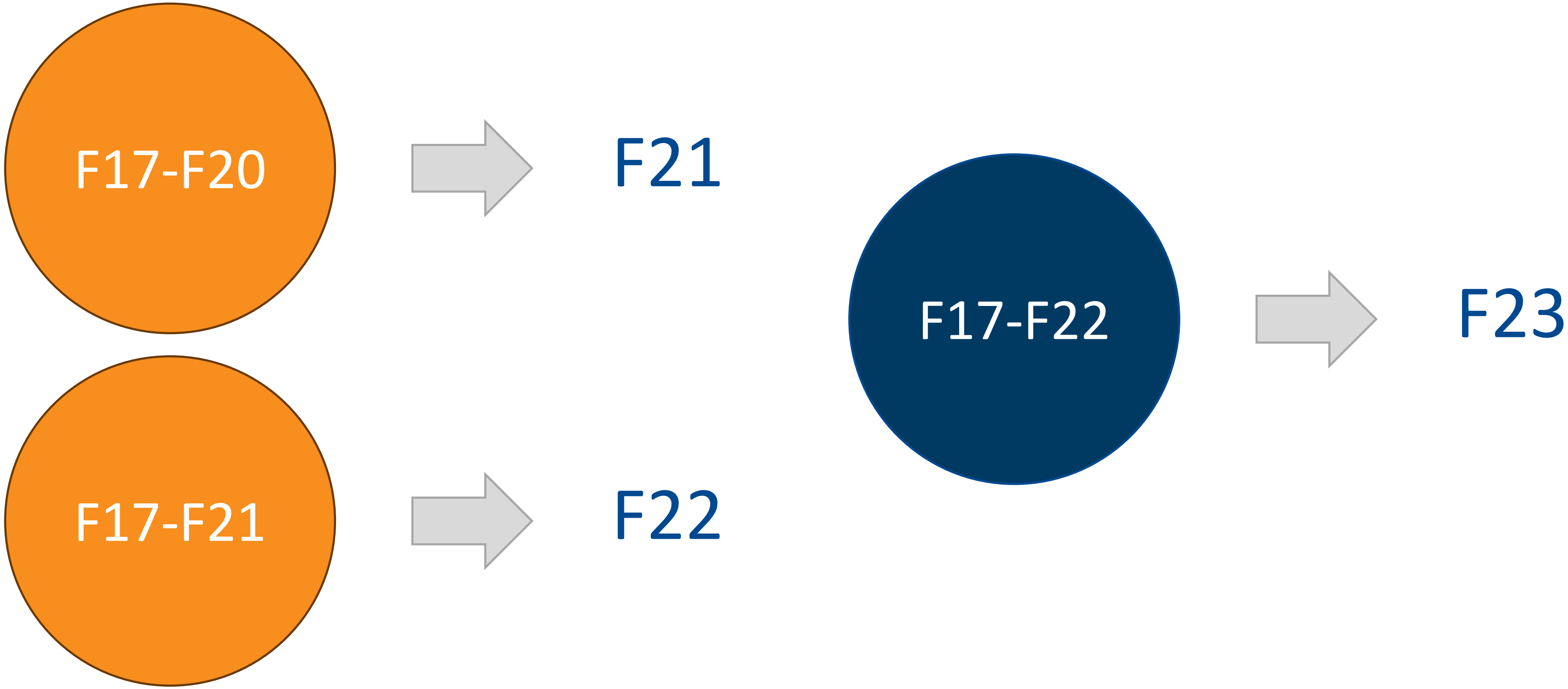
Balancing the Sample

- Approximately 75% of students retain every semester, so the model will be biased towards predicting "retained"
- Randomly oversampled the retained group to balance the sample
- 50/50 split in retained vs not retained outcomes

Missing Values

- Used “missing” indicators for GPA, test scores, AGI etc…
- Also considered imputation but found the results to be similar

Training and Testing



F17-F20

F21

F17-F21

F22

F17-F22

F23

Hyper Parameter Tuning

Parameter	Description	Value Used
Number of Trees	**Number of trees in the forest**	**1400**
Max Depth	**Max levels in a tree**	**40**
Minimum number of samples for split	Minimum number of samples required to split a node	2
Minimum number of samples for leaf	Minimum number of samples required at each leaf node	1
Bootstrap	True or False	True
Error Measure	‘gini’ or ‘entropy’	‘entropy’

Results

Variable Importance

We had surprising and not surprising results in the ranking of variable importance

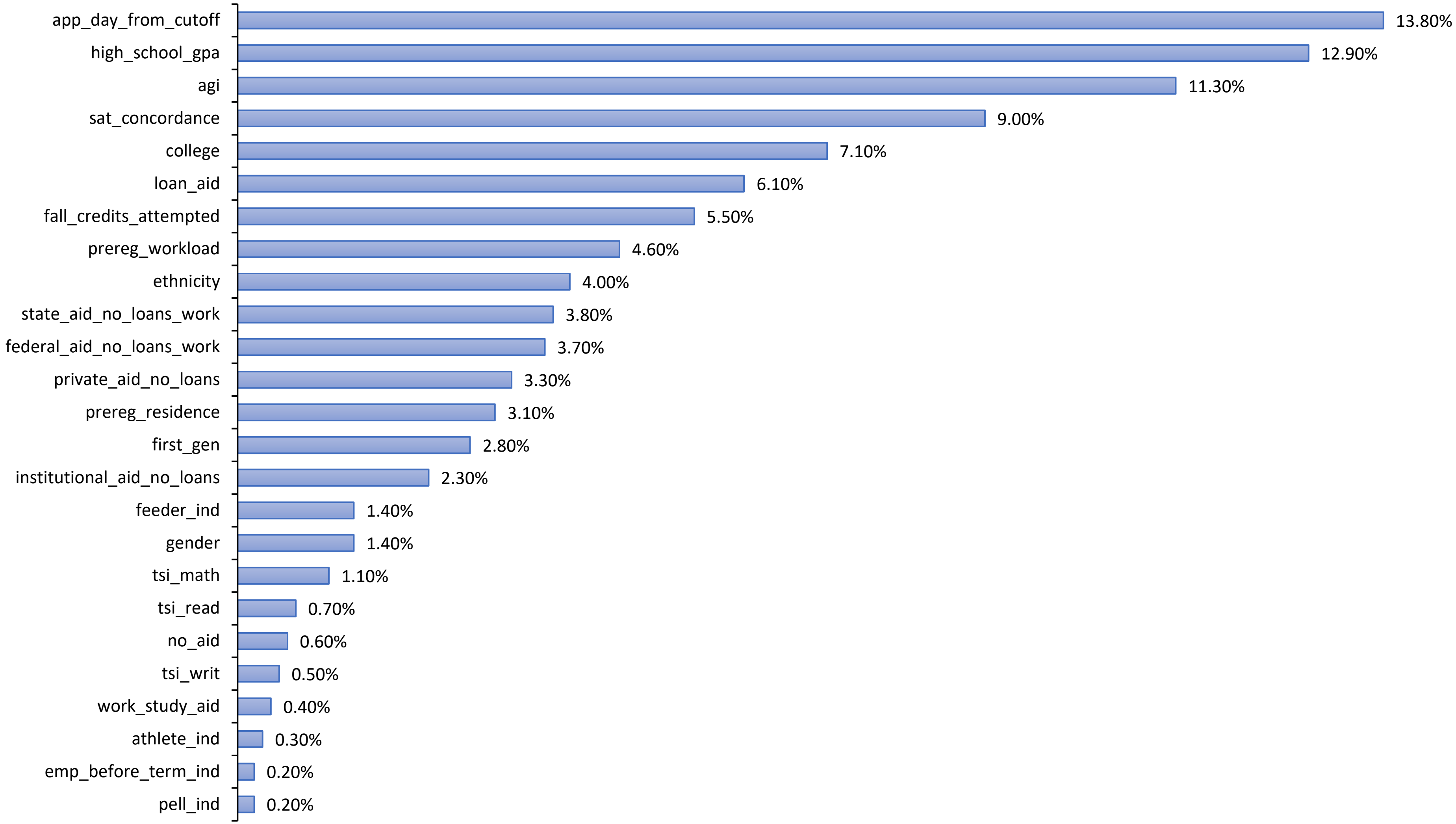
Accuracy

We see accuracy improving as more cohorts are included in the training

Grouping

We grouped students based on their predicted probability of retention

Variable Importance



Accuracy

Table 3. Confusion matrix for first train/test iteration (training Fall 2017-2020, testing Fall 2021) – Overall accuracy 71.9%

		Actual Outcome		
		Not Retained	Retained	Predicted Value
Model Prediction	Not Retained	78	112	41.0%
	Retained	705	2013	
	Model Sensitivity	9.9%		

Table 4. Confusion matrix for second train/test iteration (training Fall 2017-2021, testing Fall 2022) – Overall accuracy 72.54%

		Actual Outcome		
		Not Retained	Retained	Predicted Value
Model Prediction	Not Retained	193	278	40.9%
	Retained	654	2269	
	Model Sensitivity	22.7%		

We see evidence the model is improving as more data is added

Band Retention Probability

Band (PRV range)	Fall 2022 Band Retention Probability	Expected Student Retention Counts Fall 2023
1 (.0 ≤ PRV < .1)	Insufficient Data	--- (N=0)
2 (.1 ≤ PRV < .2)	Insufficient Data	--- (N=9)
3 (.2 ≤ PRV < .3)	46% (15/28)	24 (N=53)
4 (.3 ≤ PRV < .4)	57% (54/125)	84 (N=148)
5 (.4 ≤ PRV < .5)	62% (120/313)	242 (N=391)
6 (.5 ≤ PRV < .6)	72% (491/683)	554 (N=769)
7 (.6 ≤ PRV < .7)	73% (688/940)	700 (N=959)
8 (.7 ≤ PRV < .8)	82% (664/807)	673 (N=821)
9 (.8 ≤ PRV < .9)	84% (321/382)	319 (N=380)
10 (.9 ≤ PRV < 1)	93% (103/111)	65 (N=70)

Deployment

Purpose

To assist advisors, colleges, departments, etc. in identifying students who may need a little extra contact.

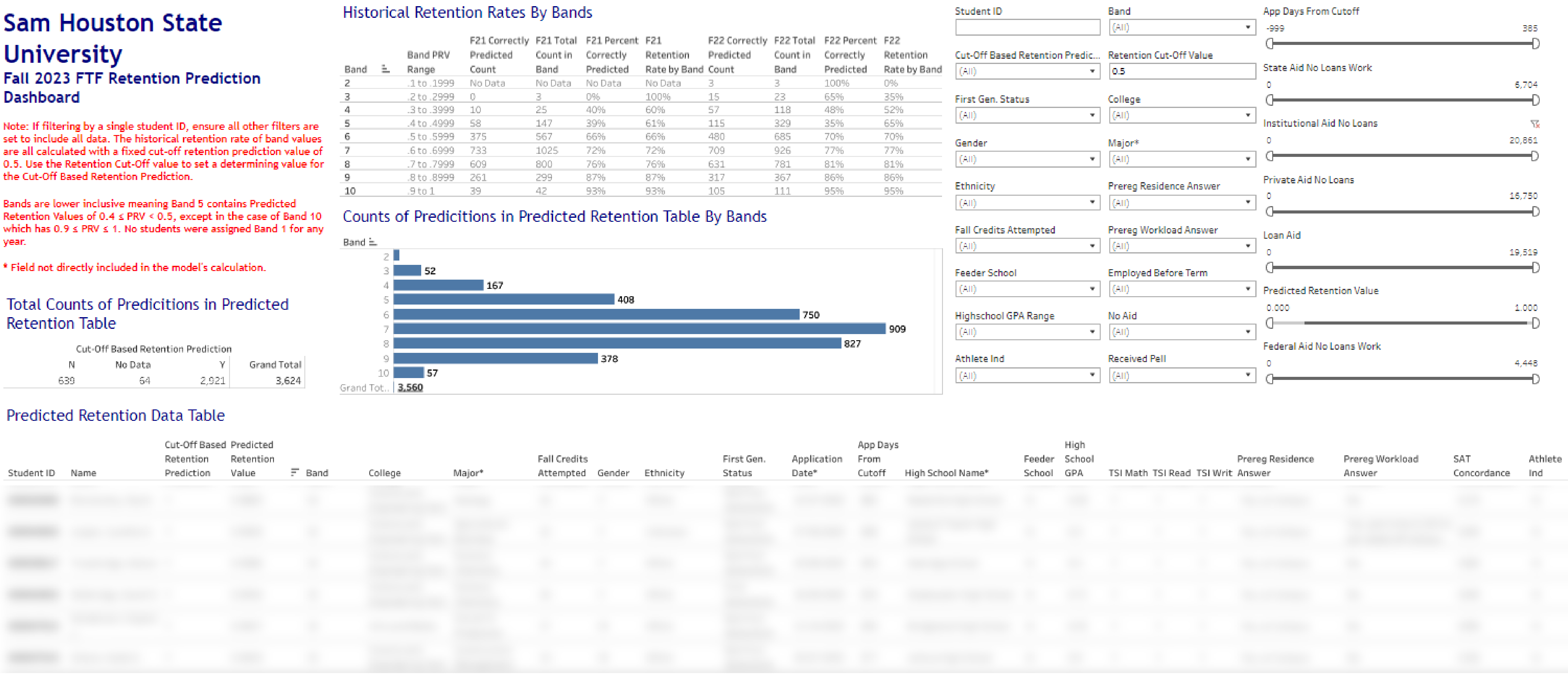
Tool

Designed to allow users to filter the FTF 2023 cohort by student details (e.g. college, department, first gen, and of course PRV score and group).

Insights from Variable Importance Analysis

Behavioral indicators and financial status seem to be the best predictors of retention.

Tableau Dashboard Tool



Sam Houston State University

Fall 2023 FTF Retention Prediction Dashboard

Note: If filtering by a single student ID, ensure all other filters are set to include all data. The historical retention rate of band values are all calculated with a fixed cut-off retention prediction value of 0.5. Use the Retention Cut-Off value to set a determining value for the Cut-Off Based Retention Prediction.

Bands are lower inclusive meaning Band 5 contains Predicted Retention Values of 0.4 ≤ PRV < 0.5, except in the case of Band 10 which has 0.9 ≤ PRV ≤ 1. No students were assigned Band 1 for any year.

* Field not directly included in the model's calculation.

Total Counts of Predicitions in Predicted Retention Table

Cut-Off Based Retention Prediction			
N	No Data	Y	Grand Total
639	64	2,921	3,624

Historical Retention Rates By Bands

Band	Band PRV Range	F21 Correctly Predicted Count	F21 Total Count in Band	F21 Percent Correctly Predicted	F21 Retention Rate by Band	F22 Correctly Predicted Count	F22 Total Count in Band	F22 Percent Correctly Predicted	F22 Retention Rate by Band
2	.1 to .1999	No Data	No Data	No Data	No Data	3	3	100%	0%
3	.2 to .2999	0	3	0%	100%	15	23	65%	35%
4	.3 to .3999	10	25	40%	60%	57	118	48%	52%
5	.4 to .4999	58	147	39%	61%	115	329	35%	65%
6	.5 to .5999	375	567	66%	66%	480	685	70%	70%
7	.6 to .6999	733	1025	72%	72%	709	926	77%	77%
8	.7 to .7999	609	800	76%	76%	631	781	81%	81%
9	.8 to .8999	261	299	87%	87%	317	367	86%	86%
10	.9 to 1	39	42	93%	93%	105	111	95%	95%

Counts of Predicitions in Predicted Retention Table By Bands

Student ID:

Band: (All)

App Days From Cutoff: -999 – 385

Cut-Off Based Retention Predic...: (All)

Retention Cut-Off Value: 0.5

State Aid No Loans Work: 0 – 6,704

First Gen. Status: (All)

College: (All)

Institutional Aid No Loans: 0 – 20,861

Gender: (All)

Major*: (All)

Private Aid No Loans: 0 – 16,750

Ethnicity: (All)

Prereg Residence Answer: (All)

Loan Aid: 0 – 19,519

Fall Credits Attempted: (All)

Prereg Workload Answer: (All)

Predicted Retention Value: 0.000 – 1.000

Feeder School: (All)

Employed Before Term: (All)

Federal Aid No Loans Work: 0 – 4,448

Highschool GPA Range: (All)

No Aid: (All)

Athlete Ind: (All)

Received Pell: (All)

Predicted Retention Data Table

Student ID	Name	Cut-Off Based Retention Prediction	Predicted Retention Value	Band	College	Major*	Fall Credits Attempted	Gender	Ethnicity	First Gen. Status	Application Date*	App Days From Cutoff	High School Name*	Feeder School	High School GPA	TSI Math	TSI Read	TSI Writ	Prereg Residence Answer	Prereg Workload Answer	SAT Concordance	Athlete Ind
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Conclusion

Reassessment and Retraining

Future plans involve reassessing and retraining model with Fall 2023 cohort data to ensure continued relevance and accuracy.

Updating the Model

There are opportunities to update the model with post-enrollment (e.g. LMS, student support services, etc.) data for enhanced accuracy.

Enrichment of the Model

The model can be enriched with additional engagement data to capture a more comprehensive picture.

References

Aljohani, O. (2016). A comprehensive review of the major studies and theoretical models of student retention in higher education. Higher education studies, 6(2), 1-18.

Baranko Faught, L. L. (2023). Efficacy of early intervention as a retention tool (Order No. 30810765). Available from ProQuest Dissertations & Theses Global. (2917722848). Retrieved from https://ezproxy.shsu.edu/login?url=https://www.proquest.com/dissertations-theses/efficacy-early-intervention-as-retention-tool/docview/2917722848/se-2

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Venit, E. (2023, February 21). How will we measure student success in the 2020’s?: A review of how student success metrics have evolved over time—and where they might go in the future. EAB. https://eab.com/insights/blogs/student-success/evolution-of-student-successmetrics/

Questions?