CALIFORNIA STATE UNIVERSITY
FULLERTON™

Graduation Predictive Model to Support Student Success and GI2025

CAIR Conference 2021

Afshin Karimi & Su Swarat
Office of Assessment & Institutional Effectiveness
California State University, Fullerton

10/28/2021

Graduation Initiative 2025

CSU system-wide initiative to raise graduation rates and to eliminate equity gaps in degree completion.

At Fullerton, our focus is mostly on increasing 4-year graduation rate of our FTF cohorts and on closing the gaps.

Example Data-Driven GI2025 Projects at CSUF

- “Caps on heads” project
- Characteristics of our 4-year graduates
- Triple-opportunity/Quadruple-opportunity student groupings
- Key first-year course failures & their relationship to student attrition (a data mining study)

Most recent GI2025 data-driven effort:
4 Year degree prediction model

Objective:

- To build a predictive data mining model (supervised learning) using historical student data in order to predict the 4-year graduation outcome of first-time freshmen early in their academic career

Objectives continued…

- Training data set includes about 13,000 students from our University fall 2013, 2014 and 2015 FTF cohorts
- Train and validate the model, then test with a future cohort (fall 16 FTF cohort) when that cohort's 4-year graduation outcome becomes available
- Apply the model on the newest FTF cohort to predict graduation outcomes, and identify students who are not likely to graduate in 4 years

First year retention is key to graduation

- Our goal is to predict students who may not graduate early before we lose some of them due to attrition

- From our most recent 3 freshman cohort aggregate:
  - 11.4% attrition between 1st and 2nd year
  - 6.7% attrition between 2nd and 3rd year
  - 3.7% attrition between 3rd and 4th year

Timing

- Predict early and start the interventions during students' first year

- How early?
  - Ideally before/upon matriculation
    - Built and tested a model using only students' pre-college variables, but the model's prediction accuracy was low and not acceptable
  - After completion of first semester (early January)
    - First semester grades and units as well as second semester's registered units are available and used in the model
    - Model yielded satisfactory prediction accuracy

Algorithms

- 2 different classification algorithms to build two models:
  - Binary Decision Tree
  - Logistic Regression

- Train and build both models in parallel, and upon completion of the validation phase, pick a model with higher accuracy

The Predictor (independent) Variables

- Ran several iterations of Logistic Regression (while adding/removing independent variables in the equation iteratively while observing the resulting P-values and coefficients) in order to pick the best combination of independent variables.

- Calculated the information gain ratio of all the predictor variables. Variables with minimum entropy would yield highest information gain

Final List of Predictor Variables (Fall 2020 model)

Variable	Type	Range
Sex	binomial	M/F
URM	binomial	Y/N
Pell Recipient	binomial	Y/N
First Gen. to Attend College	binomial	Y/N
EI Group	integer	Group 0 thru 8
College Group	binomial	Hierarchical/Non-Hierarchical
Units Earned-Fall	integer	0 -22
A2 GE Course Grade Group	polynomial	B- or better, C- to C+, D+ or lower, No attempt
Fall semester GPA	real	0.0 to 4.0
Spring semester units registered	integer	0 - 21
Early NSO attended	binomial	Y/N

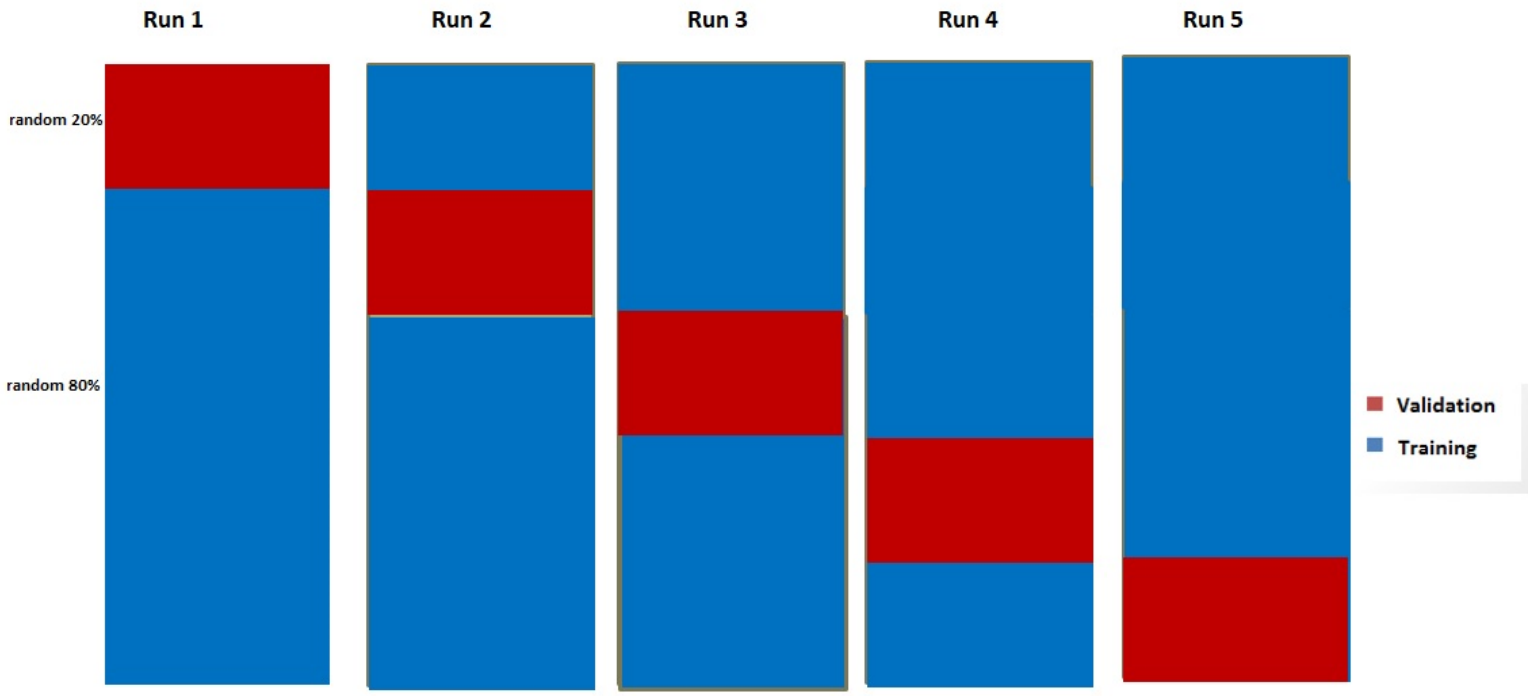
Imbalanced Data

- Our training dataset is somewhat imbalanced. About 25% of the records graduated in 4 years (minority class) and about 75% did not (majority class)
- Many classification models' performance are biased towards the majority class
- To resolve this issue, the minority class data in the training dataset is oversampled

(5 fold) Cross Validation

- Random 20% of the training dataset was set aside for validation; trained the model using the other 80% and validated it with the set-aside 20%
- Repeated the above 4 more times, each time setting aside a different random 20%
- The model validation (the generated prediction accuracy) is the average of the 5 runs

(5 fold) Cross Validation Continued…



(5 fold) Cross Validation Results

Binary Decision Tree

		Predicted 4 Year Degree		
		Yes	No	Total
Actual 4 Yr. Degree	Yes	537	433	970
	No	931	2516	3447

Overall Prediction Accuracy=(537+2516)/3447 = 69.1%

Logistic Regression

		Predicted 4 Year Degree		
		Yes	No	Total
Actual 4 Yr. Degree	Yes	591	379	970
	No	752	2685	3447

Overall Prediction Accuracy=(591+2685)/3447 = 74.4%

Logistic Regression method yielded higher overall prediction accuracy

Model Testing: Fall 16 FTF Cohort

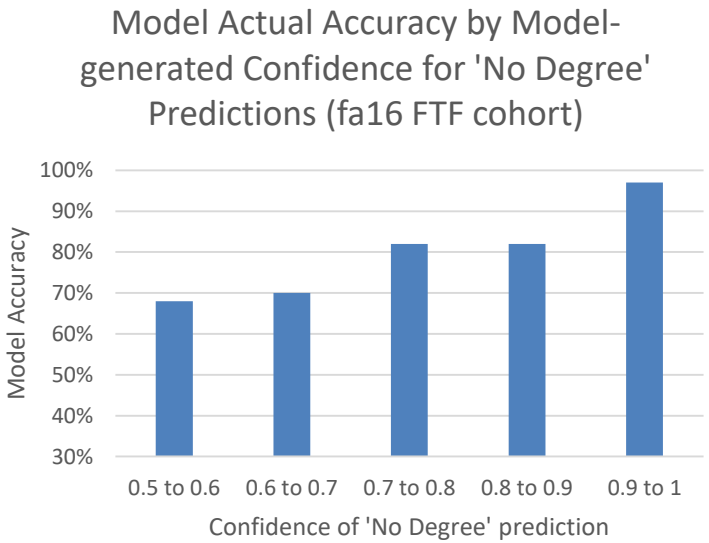
Tested the model with the fall 16 FTF cohort after the summer 2020 degrees were posted

		Predicted 4 year degree		
		No	Yes	Total
Actual 4 year degree	No	2279	577	2856
	Yes	602	816	1418
Overall Prediction Accuracy=(2279+816)/4274=72.4%				

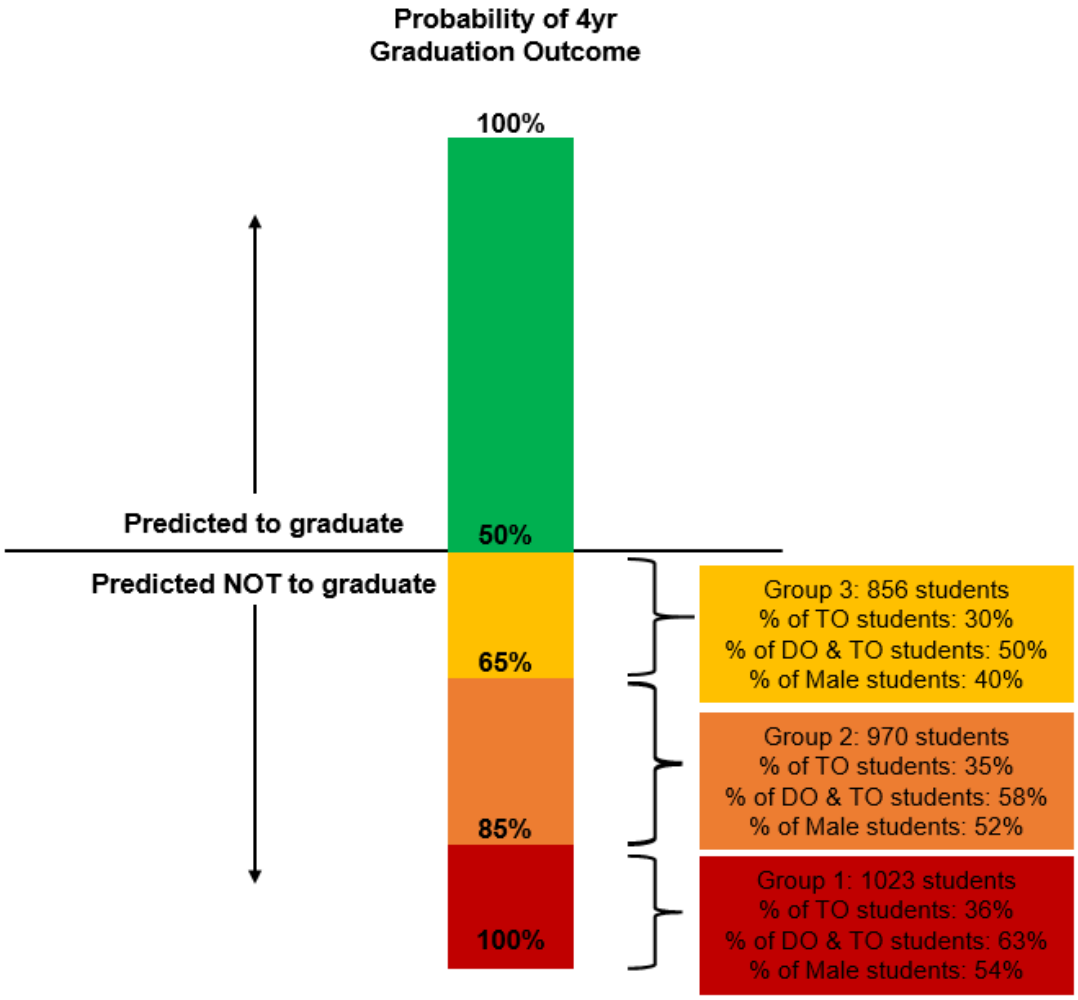
Confidence of Prediction

The model outputs (per student):

- 4 yr. degree prediction (Yes/No)
- Confidence of that prediction (between 50% to 100%)



Model Accuracy and % Triple-opportunity



How is the model used

- List of students and their predicted 4-year graduation outcome (and probability) are distributed to the colleges, highlighting students who could benefit from concentrated support
- Couple the list with triple-opportunity information to zoom in on students who the university should focus on within the GI2025 context
- Freshmen grant program implemented in summer 2021 based on the list
- Sophomore ”bridge” program re-designed by college of H&SS based on the list

Year 2 enhancements

- Re-train the model using fall 14, 15, 16 freshman cohorts combined:
  - Give a bump to students in the following majors who either had the credit for or passed Calculus 1 in their first term: Math, Physics, all majors in the college of Engineering & Computer Science
  - Training data set particularly imbalanced for 3 colleges (Arts, Engineering & Computer Science and Natural Sciences & Mathematics). Provide additional oversampling for records in those colleges
  - Add a bump to College of the Arts students who passed 5 or more Fine Arts classes in high school

Year 2 enhancements, continued…

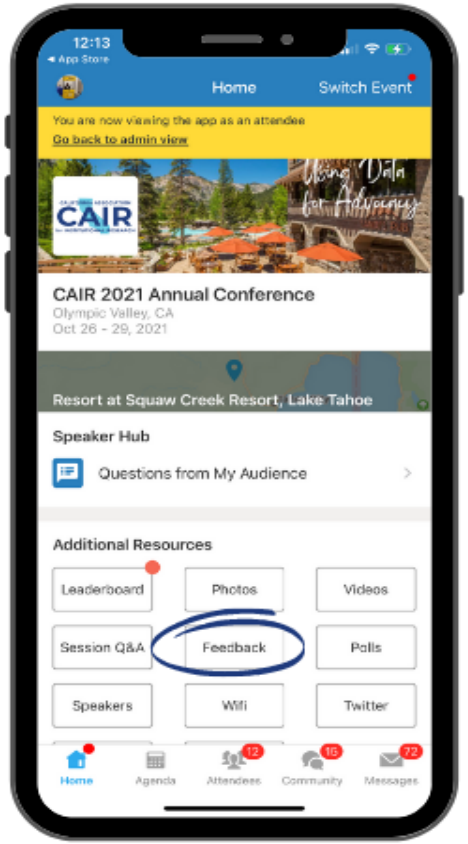
- In fall 2021 when fall 17 cohort's 4 year degree outcomes become available, test the model using that cohort

- Run the model in January 2022 on the incoming fall 2021 freshman cohort (after they complete 1 semester at CSUF)

Questions:

data@fullerton.edu

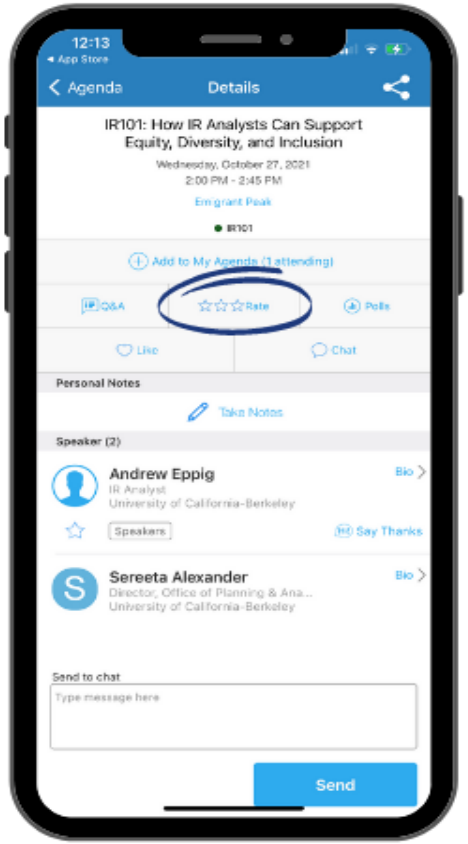
Reminder: Complete your session evaluations



OPTION 01

Home - Feedback

- Navigate to the **Home** page
- Click on **Feedback**
- Select **Session Feedback**
- Select the name of the session that you attended



OPTION 02

Agenda - Session

- Navigate to **Agenda** on the bottom menu
- Select session name
- Click ☆☆☆ **Rate**