Towards Equity and Algorithmic Fairness in Student Grade Prediction

Weijie Jiang*, Zachary A. Pardos* {jiangwj, pardos}@berkeley.edu
University of California, Berkeley
(* Both authors contributed equally to this research.)

CAHL
Computational Approaches to Human Learning (CAHL) research lab

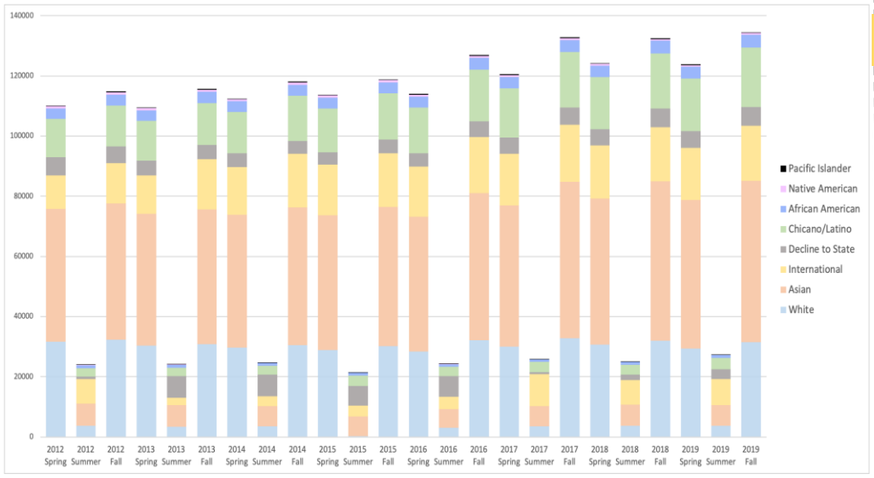
1. Goal & Datasets

Goals:

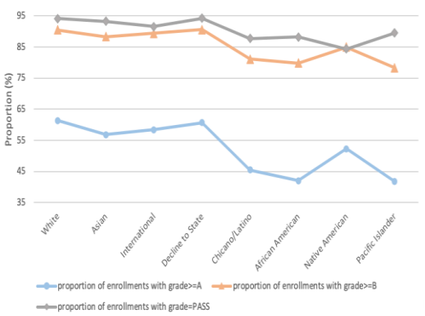
- With **FAIRNESS** as the aim, trial several strategies for both label and instance balancing to minimize differences in algorithm performance with respect to race.
- With **EQUITY** of educational outcome as the aim, trial strategies for boosting predictive performance on historically underserved groups and find success in sampling those groups in inverse proportion to their historic outcomes.

Datasets:

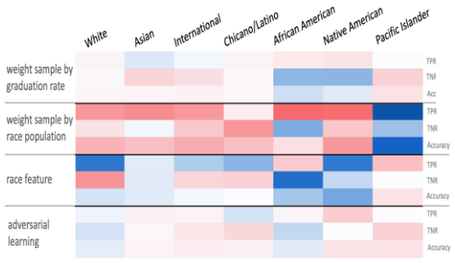
- **Student enrollment data:** anonymized student course enrollments from Spring 2012 through Fall 2019 of 82,309 undergraduates with a total of 1.97 million enrollments. Grade types include letter grades (i.e., A, B, C, D, F) with some courses allowing students to elect to be graded based on a PASS/No-PASS score.
- **Student Demographic Data** : gender, race, entry status, and parents income when admitted.



Distribution of enrollments across semesters by race

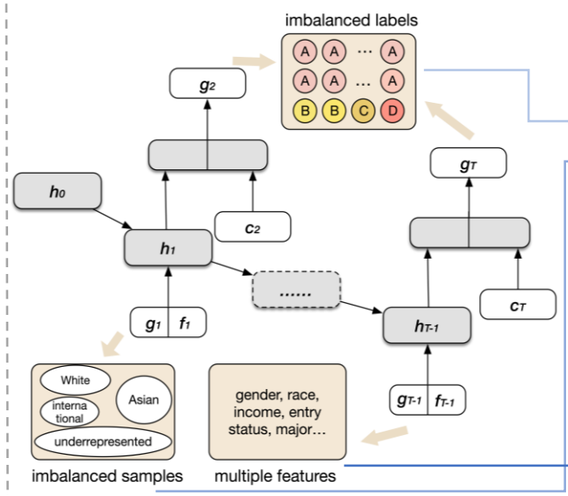


Grade distribution by race



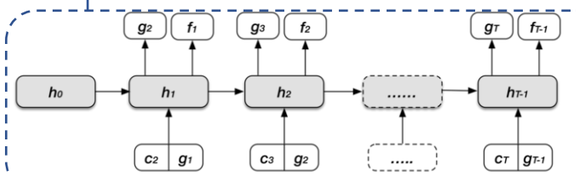
Heat map of performance of the four fairness and equity-based strategies.
White = same as baseline (no strategy)
Blue = improvement over baseline
Red = reduction compared to baseline

2. Strategies to Mitigate Bias in Grade Prediction



Pre-processing strategies to improve fairness

Strategy	Name	Stage
fairness through unawareness	default(loss)	-
weight loss by grade label	grade label weighted loss	data construction
weight loss by sample	alone, grad-rate (wgh), equal (wgh)	data construction
sensitive feature added to input	race (feature)	data construction
multiple features added to input	multi	data construction
remove features for prediction	infer-rmv	inference (prediction)
adversarial learning	adversarial	model training



adversarial loss: $Loss_F = -\sum_t \hat{f}_t^T \log f_t$

$$L = -\sum_t \sum_{i, g^i_{t+1} \neq 0} (\hat{g^{i1}}^T_{t+1} \log g^{i1}_{t+1} + \hat{g^{i2}}^T_{t+1} \log g^{i2}_{t+1}) + Loss_F$$

3. Experiment Results Analysis

		White	Asian	International	Chicano/Latino	African American	Native American	Pacific Islander	Overall	Range	STD
TPR(%)	default	80.10	79.67	78.16	70.31	**72.46**	78.34	72.58	78.39	9.79	4.02
	grad-rate(wgh)	79.89	**80.07**	78.27	70.09	71.96	77.71	72.58	**79.82**	9.98	4.13
	equal(wgh)	77.36	76.65	75.49	69.93	68.51	74.52	**79.03**	79.46	10.52	3.90
	race(feature)	**82.70**	79.99	**79.10**	**71.72**	71.17	**80.89**	70.97	79.53	11.73	5.14
	adversarial	80.27	79.37	77.91	70.79	72.26	77.07	72.58	78.42	**9.48**	**3.80**
TNR(%)	default	70.76	74.76	**73.56**	**81.01**	78.63	77.62	80.23	**74.91**	10.25	3.75
	grad-rate(wgh)	70.67	73.68	72.79	80.92	79.99	**79.02**	79.07	73.89	10.25	4.09
	equal(wgh)	70.04	74.89	72.17	78.27	80.20	76.22	**81.40**	73.69	11.36	4.15
	race(feature)	67.95	**75.09**	72.53	79.84	**81.42**	78.32	80.23	74.21	13.47	4.89
	adversarial	**71.27**	74.61	72.99	80.03	79.34	77.62	79.07	74.75	**8.76**	**3.45**
Accuracy(%)	default	76.50	77.55	76.25	76.14	76.04	78.00	77.03	76.86	1.96	0.76
	grad-rate(wgh)	76.33	77.31	75.99	76.00	76.62	78.33	76.35	76.82	2.34	0.85
	equal(wgh)	74.54	75.89	74.11	74.48	75.29	75.33	**80.41**	76.93	6.30	2.16
	race(feature)	**77.01**	**77.88**	**76.36**	**76.15**	**77.11**	**79.67**	76.35	**77.19**	3.52	1.23
	adversarial	76.80	77.31	75.86	75.83	76.37	77.33	76.35	76.81	**1.50**	**0.62**

- The **adversarial learning** strategy achieved all the minimums of range and standard deviation for TPR, TNR, and accuracy, demonstrating the best group fairness among all the compared strategies.
- **Presenting race explicitly to the input** of the model led to the most unfair results out of all strategies, though also the most accurate, overall.

- **Weighting the loss function by grade label** boosted accuracy for Chicano/Latino, African American, Native American, and Pacific Islander students without sacrificing much accuracy for White, Asian, and International students.
- The **equity of outcome approach**, which sampled instances by group with inverse proportion to a historic educational outcome (grad-rate), was effective in boosting the predictive accuracy of most of the historically underserved groups, and increase the TNR and accuracy for African American and Native American students, who have recorded the lowest on-time graduation rates.

Fourth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, A virtual conference, May 19-21, 2021