Towards Equity and Algorithmic Fairness in Student Grade Prediction

By Weijie Jiang and Zachary A. Pardos

Contribution

- Proposing data processing approaches to remove biases for grade prediction tasks
- Adapting adversarial learning architecture to remove bias for grade prediction
- Implementing a testing for multiple

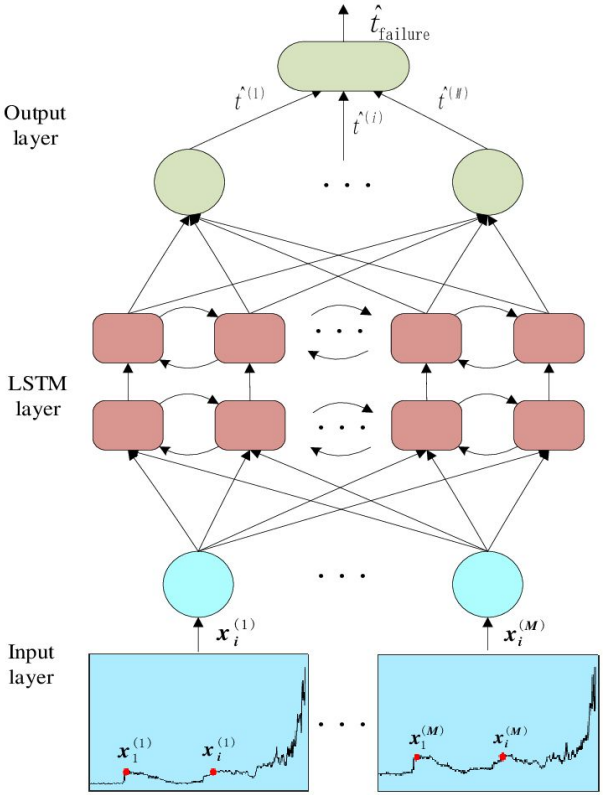
Motivation

- Grade prediction is an important task to help struggling students and college admission
- Equalized odds and equalized opportunities
- We may need a more fair but less performant approach

Methods:Base Model

$$L_{masked} = MaskedCrossEntropy(\hat{\boldsymbol{g}}_{t+1}, \boldsymbol{g}_{t+1})$$

$$= -\sum_{t} \sum_{i, \hat{g}^{i}_{t+1} \neq 0} (\hat{\boldsymbol{g}^{i1}}^{T}_{t+1} \log \boldsymbol{g}^{i1}_{t+1} + \hat{\boldsymbol{g}^{i2}}^{T}_{t+1} \log \boldsymbol{g}^{i2}_{t+1})$$



Methods: Data construction

Sensitive attribute balancing

$$L_{wgs} = -\sum_{t} \lambda(r(\hat{\boldsymbol{g}}_{t+1}))\, MaskedCrossEntropy(\hat{\boldsymbol{g}}_{t+1}, \boldsymbol{g}_{t+1})$$

$$\lambda(\boldsymbol{r}) = 1/\boldsymbol{r}, \qquad \lambda(\boldsymbol{d}) = 1 - \boldsymbol{d}$$

Grade label balancing

$$L_{wbl} = -\sum_{t} \sum_{i, \hat{\boldsymbol{g}}^{i}_{t+1} \neq 0} \sigma(\hat{\boldsymbol{g}}^{i}_{t+1}) (\hat{\boldsymbol{g}^{i1}}^{T}_{t+1} \log \boldsymbol{g}^{i1}_{t+1} + \hat{\boldsymbol{g}^{i2}}^{T}_{t+1} \log \boldsymbol{g}^{i2}_{t+1})$$

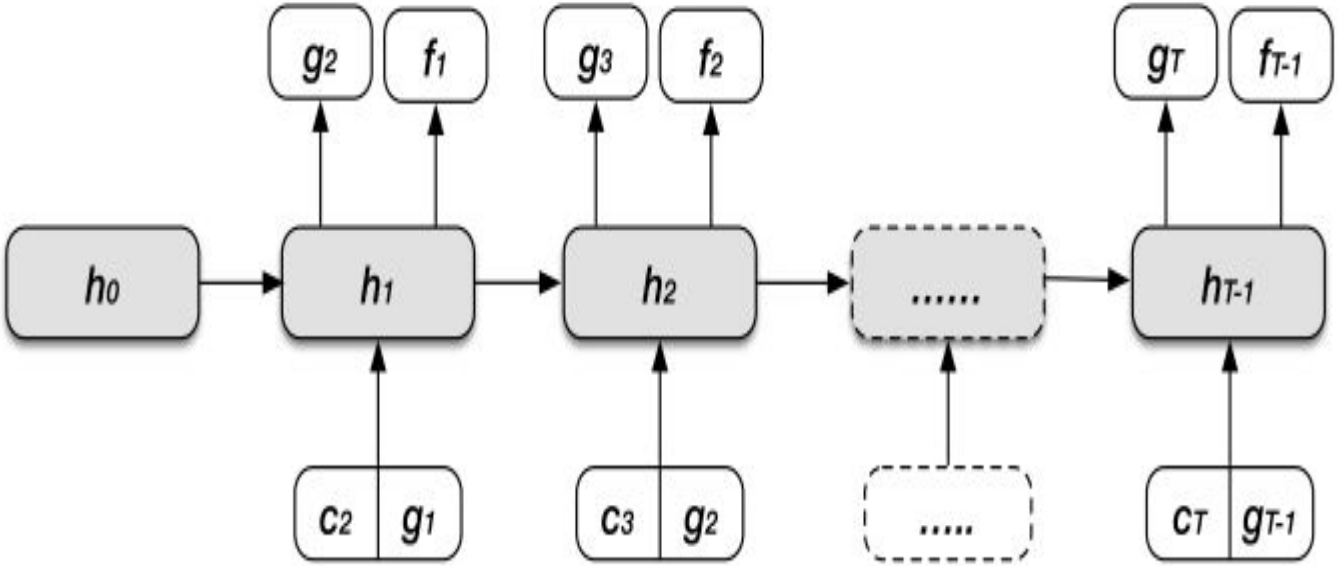
$$\sigma(\hat{\boldsymbol{g}}^{i}_{t+1}) = \frac{1/p(\hat{\boldsymbol{g}}^{i}_{t+1})}{\sum_{t} 1/p(\hat{\boldsymbol{g}}^{i}_{t+1})}$$

Final Loss

$$L_{wbsl} = -\sum_{t} \lambda(r(\hat{\boldsymbol{g}}_{t+1})) \sum_{i, \hat{\boldsymbol{g}}^{i}_{t+1} \neq 0} \sigma(\hat{\boldsymbol{g}}^{i}_{t+1}) (\hat{\boldsymbol{g}^{i1}}^{T}_{t+1} \log \boldsymbol{g}^{i1}_{t+1} + \hat{\boldsymbol{g}^{i2}}^{T}_{t+1} \log \boldsymbol{g}^{i2}_{t+1})$$

Methods: Adversarial learning

Model does not learn hidden states that can be used for discrimination



Methods: Adversarial learning

Adversarial loss

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

Total Loss

$$L = -\sum_t \sum_{i,\hat{g}_{t+1}^i \neq 0} (\hat{g^{i1}}_{t+1}^T \log g_{t+1}^{i1} + \hat{g^{i2}}_{t+1}^T \log g_{t+1}^{i2}) + \alpha \sum_t \hat{f}_t^T \log f_t$$

Methods: Sensitive feature inclusion

- Default not additional attributes
- Leave out sensitive features in inference
- Include race attribute
- Multiple sensitive features

Experiments: Dataset

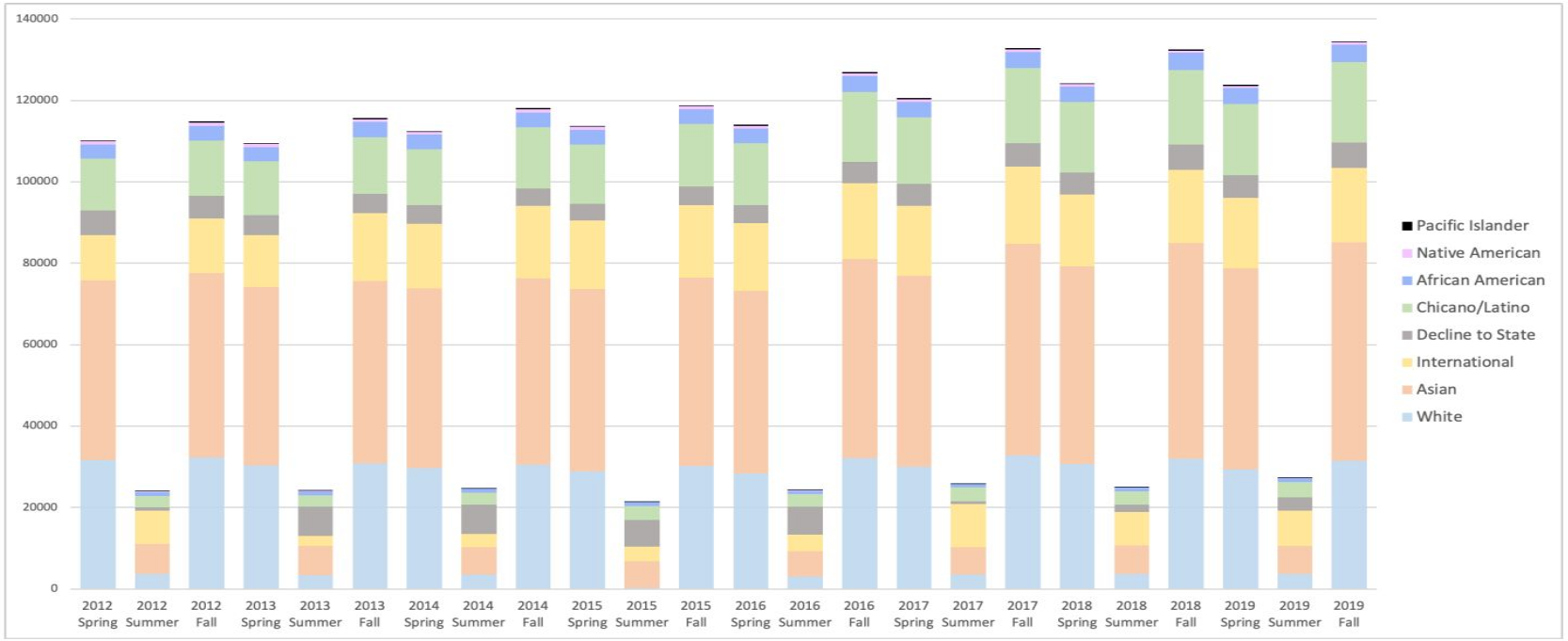
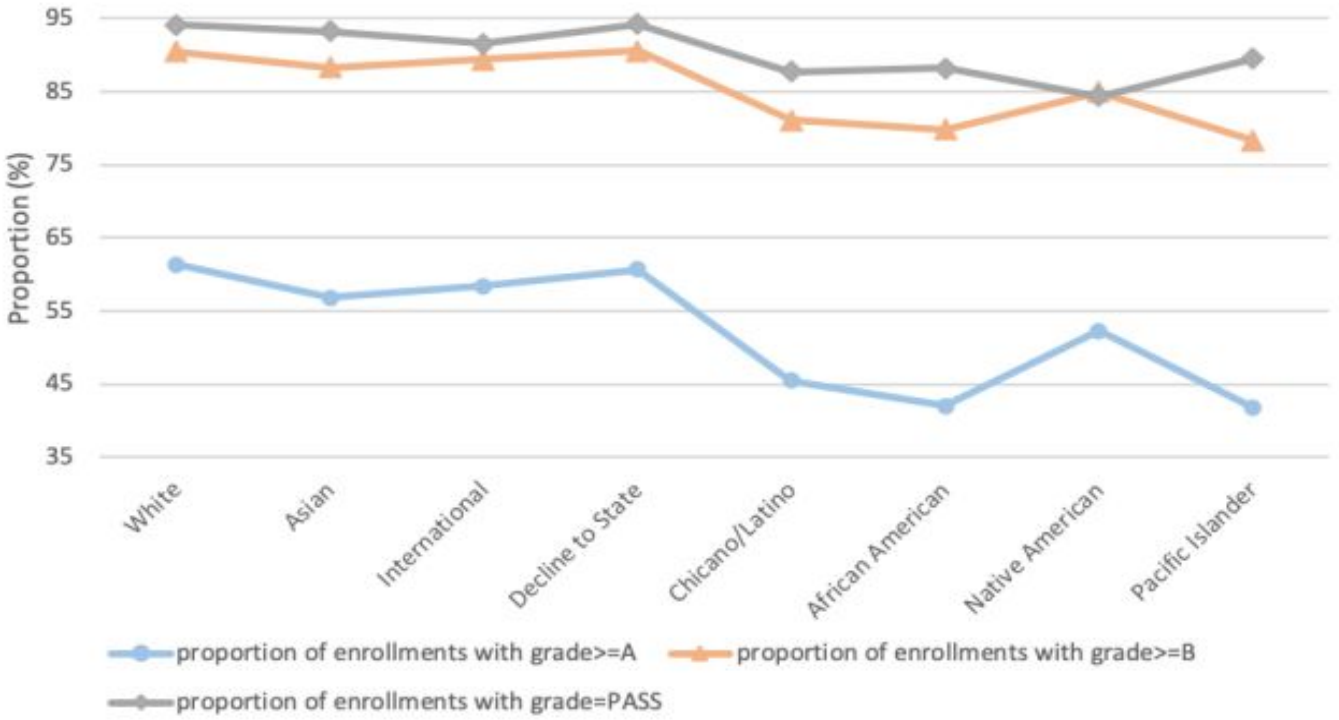


Figure 1: Distribution of enrollments by race across semesters

Experiments: Dataset



Experiments: Metrics

- True positive rate (opportunities for performing students)
- True negative rate (support for struggling students)
- Accuracy (Overall performance)
- Range of other metrics among subgroups (Fairness)
- Standard deviation of other metrics among subgroups (Fairness)

Results: Label balancing

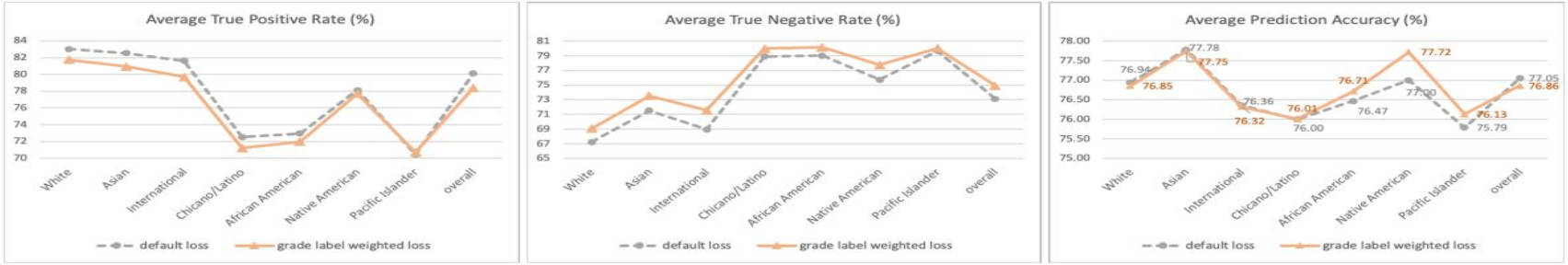


Figure 5: Results of comparison between models with unweighted loss and models with weighted loss by grade label

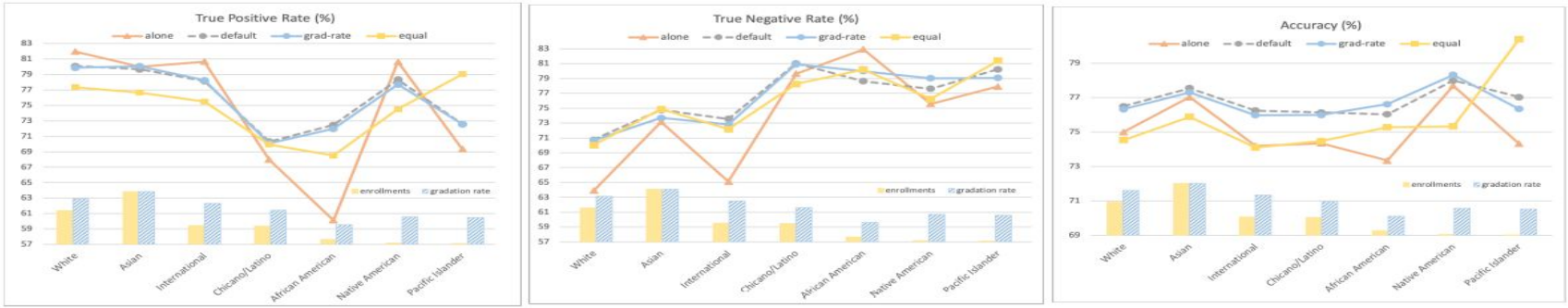


Figure 6: Evaluation results comparison between models with weighted loss by race

Results: Adding sensitive attribute

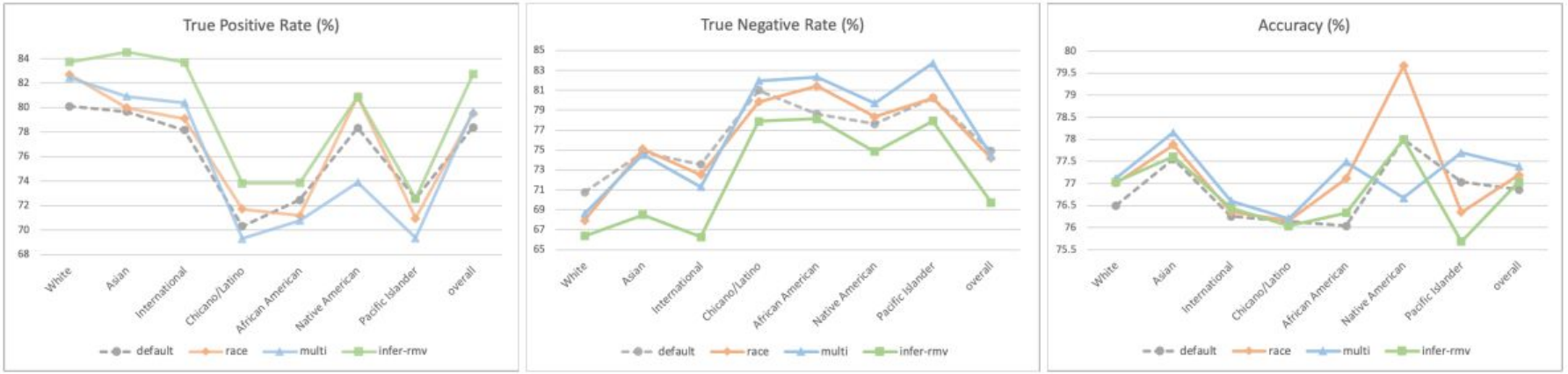


Figure 7: Results of models adding sensitive student attributes to the input

Results: Fairness results

		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**

Conclusion

- Adding the race feature is the least fair (it just had the least drop?)
- Adversarial learning is the most fair (is it significant?)
- Equality of outcome approach boosted the accuracy (How true is this?)

Questions?