Beyond Predictive Accuracy: Fairness and Bias in Predicting Test Anxiety SUPPLEMENTARY

Anonymous Author(s)

Anonymous Institute

Table 1: Survey Questions and their aliases. *Hereinafter, we will refer to specific survey items by their aliases.MSR items (first twelve rows) are highlighted in green and test anxiety items (last five rows) are highlighted in gray

Survey Item	Alias*
1: During class time I often miss important points because I'm thinking of other things.	`distracted_during_class`
2: When reading for this course, I make up questions to help focus my reading.	`formulate_guiding_questions`
3: When I become confused about something I'm reading for this class, I go back and try to figure it out.	`clarify_confusing_content`
4: If course readings are difficult to understand, I change the way I read the material.	`adjust_reading_strategy`
5: Before I study new course material thoroughly, I often skim it to see how it is organized.	`preview_course_material`
6: I ask myself questions to make sure I understand the material I have been studying in this class.	`self_check_understanding`
7: I try to change the way I study in order to fit the course requirements and the instructor's teaching style.	`adapt_study_methods`
8: I often find that I have been reading for this class but don't know what it was all about.	`mindless_class_reading`
9: I try to think through a topic and decide what I am supposed to learn from it rather than just reading it over.	`identify_learning_objectives`
10: When studying for this course I try to determine which concepts I don't understand well.	`assess_concept_mastery`
11: When I study for this class, I set goals for myself in order to direct my activities in each study period.	`set_study_goals`
12: If I get confused taking notes in class, I make sure I sort it out afterwards.	`review_unclear_notes`
1: When I take a test, I think about how poorly I am doing compared with other students.	`comparison_with_others`
2: When I take a test, I think about items on other parts of the test I can't answer.	`difficult_questions_fixation`
3: When I take tests, I think of the consequences of failing.	`fear_of_failure`
4: I have an uneasy, upset feeling when I take an exam.	`exam-induced_uneasiness`
5: I feel my heart beating fast when I take an exam.	`exam-induced_heart_racing`

Table 2: Factor Loadings showing the correlation (coefficient) of each of the 5 survey items to the test anxiety. CWO: `comparison_with_others`, DQF: `difficult_questions_fixation`, FOF: `fear_of_failure`, EIU: `exam-induced_uneasiness`, EHR: `exam-induced_heart_racing`.

CWO	DQF	FOF	EIU	EHR
0.59	0.54	**0.81**	0.72	0.70

Table 3: Average Predictive Accuracy of all models for threshold $\tau = 0.4$. The values are the average ± standard deviation. Boldened and red scores are the highest and least overall averages, respectively.

Model	AUC-ROC	AUC-PR	Accuracy	F1 Score	Overall Average
LR	0.70 ± 0.06	0.84 ± 0.04	0.79 ± 0.03	0.87 ± 0.02	**0.80 ± 0.04**
MLP	0.61 ± 0.06	0.82 ± 0.05	0.67 ± 0.04	0.78 ± 0.03	0.72 ± 0.04
RF	0.66 ± 0.05	0.83 ± 0.04	0.78 ± 0.04	0.87 ± 0.03	0.79 ± 0.04
SVM	0.71 ± 0.06	0.85 ± 0.04	0.75 ± 0.04	0.86 ± 0.02	0.79 ± 0.04
XGB	0.61 ± 0.07	0.80 ± 0.05	0.75 ± 0.04	0.84 ± 0.03	0.75 ± 0.05

Table 4: Threshold $\tau = 0.4$ fairness results. Predictive accuracies across Race, Sex, and Migrant Status (MS) groups, highlighting mean disparities, standard deviations, and statistical significance.

	Model	AUC-PR	AUC-ROC	Accuracy	F1 Score
Race	LR	-0.19 ± 0.06 ***	-0.28 ± 0.08 ***	-0.12 ± 0.06 ***	-0.07 ± 0.04 ***
	MLP	-0.11 ± 0.06 ***	-0.10 ± 0.08 ***	0.00 ± 0.06	0.00 ± 0.04
	RF	-0.06 ± 0.06 ***	-0.13 ± 0.09 ***	-0.01 ± 0.06	0.00 ± 0.04
	SVM	0.00 ± 0.06	-0.04 ± 0.09 **	-0.02 ± 0.06 **	-0.01 ± 0.04 **
	XGB	-0.09 ± 0.06 ***	-0.11 ± 0.09 ***	-0.04 ± 0.06 ***	-0.02 ± 0.04 ***
Sex	LR	0.07 ± 0.06 ***	0.10 ± 0.09 ***	0.04 ± 0.05 ***	0.04 ± 0.03 ***
	MLP	0.05 ± 0.06 ***	0.10 ± 0.08 ***	0.14 ± 0.06 ***	0.10 ± 0.05 ***
	RF	-0.01 ± 0.06	0.03 ± 0.08 **	0.03 ± 0.05 ***	0.02 ± 0.03 ***
	SVM	-0.03 ± 0.06 ***	-0.04 ± 0.08 **	-0.01 ± 0.05	0.00 ± 0.03
	XGB	0.04 ± 0.06 ***	0.05 ± 0.08 ***	0.01 ± 0.05	0.02 ± 0.04 **
MS	LR	0.14 ± 0.04 ***	0.22 ± 0.08 ***	0.07 ± 0.06 ***	0.05 ± 0.04 ***
	MLP	0.15 ± 0.05 ***	0.25 ± 0.08 ***	0.08 ± 0.06 ***	0.05 ± 0.05 ***
	RF	0.15 ± 0.04 ***	0.26 ± 0.08 ***	0.08 ± 0.06 ***	0.05 ± 0.04 ***
	SVM	0.12 ± 0.04 ***	0.21 ± 0.08 ***	0.01 ± 0.06	0.00 ± 0.04
	XGB	0.10 ± 0.06 ***	0.15 ± 0.12 ***	0.14 ± 0.06 ***	0.09 ± 0.04 ***

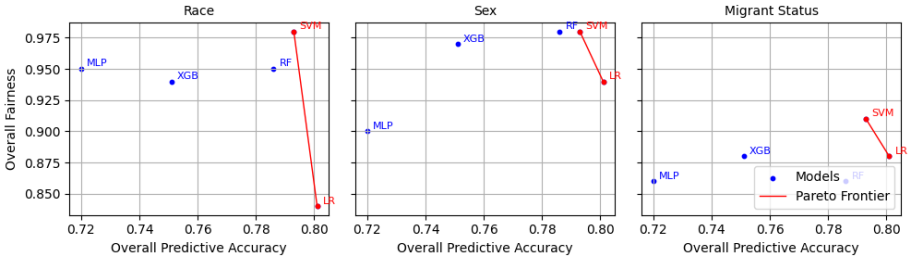

Fig. 1: Pareto frontier illustrating the trade-off between fairness and predictive accuracy. Red and blue points mark optimal and suboptimal models respectively for $\tau = 0.4$

Table 5: Threshold $\tau = 0.4$. RFG for various predictive accuracy metrics: PR(AUC-PR), ROC (AUC-ROC), Acc(Accuracy), and F1(F1 Score). Column initials represent demographic intersections: NWM (Non-White Migrant), WM (White Migrant), NWNM (Non-White Non-Migrant), WNM (White Non-Migrant), NWF (Non-White Female), NWM (Non-White Male), WF (White Female), WM (White Male), FM (Female Migrant), MM (Male Migrant), FNM (Female Non-Migrant), and MNM (Male Non-Migrant)

		Race-Migrant Status				Race-Sex				Migrant-Sex			
		NWM	WM	NWNM	WNM	NWF	NWM	WF	WM	FM	MM	FNM	MNM
LR	AUC-PR	0.10	0.04	-0.07	0.05	0.01	-0.07	0.04	0.08	0.02	0.10	0.02	-0.02
	AUC-ROC	0.17	0.08	-0.14	0.07	-0.02	-0.14	0.08	0.11	0.04	0.19	0.03	-0.04
	Accuracy	0.03	0.05	-0.05	0.03	0.03	-0.12	-0.00	0.06	-0.05	0.10	0.02	-0.03
	F1 Score	0.01	0.03	-0.03	0.02	0.02	-0.09	-0.00	0.04	-0.04	0.07	0.01	-0.03
MLP	AUC-PR	0.11	0.03	-0.06	0.03	0.02	-0.05	0.02	0.05	0.05	0.09	0.01	-0.02
	AUC-ROC	0.19	0.01	-0.10	0.02	-0.01	-0.04	0.07	0.02	0.12	0.13	0.01	-0.03
	Accuracy	0.09	-0.09	-0.04	0.01	0.02	-0.04	0.05	-0.01	0.05	-0.01	0.04	-0.03
	F1 Score	0.06	-0.08	-0.03	0.01	0.01	-0.03	0.03	-0.01	0.02	-0.01	0.03	-0.02
RF	AUC-PR	0.06	0.07	-0.05	0.02	0.03	-0.05	-0.02	0.09	0.01	0.09	-0.01	-0.01
	AUC-ROC	0.11	0.13	-0.09	0.02	0.02	-0.08	-0.02	0.12	0.03	0.19	0.02	-0.04
	Accuracy	-0.00	0.07	-0.02	-0.00	0.01	-0.03	-0.00	0.03	-0.04	0.10	0.01	-0.02
	F1 Score	-0.01	0.04	-0.01	-0.00	0.01	-0.03	-0.01	0.02	-0.04	0.06	0.01	-0.01
SVM	AUC-PR	0.03	0.08	-0.01	0.00	0.05	-0.02	-0.03	0.07	0.01	0.07	-0.01	0.01
	AUC-ROC	0.06	0.13	-0.05	0.00	0.01	-0.04	-0.02	0.09	0.02	0.15	-0.01	-0.01
	Accuracy	-0.03	0.04	-0.01	0.00	0.02	-0.04	-0.03	0.04	-0.07	0.05	-0.00	-0.00
	F1 Score	-0.02	0.03	-0.00	0.00	0.01	-0.03	-0.02	0.03	-0.05	0.03	-0.00	-0.00
XGB	AUC-PR	0.07	0.05	-0.03	0.02	0.01	-0.03	0.02	0.06	-0.05	0.15	0.04	-0.02
	AUC-ROC	0.15	-0.00	-0.06	0.03	-0.05	-0.02	0.06	0.05	-0.12	0.34	0.07	-0.06
	Accuracy	0.12	-0.01	-0.05	0.02	-0.01	-0.03	0.00	0.03	-0.13	0.20	0.02	-0.03
	F1 Score	0.07	-0.01	-0.03	0.01	-0.00	-0.03	-0.01	0.02	-0.10	0.13	0.02	-0.02

Table 6: Average Predictive Accuracy of all models for threshold $\tau = 0.6$. The values are the average ± standard deviation. Boldened and red scores are the highest and least overall averages, respectively.

Model	AUC-ROC	AUC-PR	Accuracy	F1 Score	Overall Average
LR	0.75 ± 0.05	0.74 ± 0.06	0.70 ± 0.04	0.67 ± 0.06	**0.71 ± 0.05**
MLP	0.57 ± 0.05	0.58 ± 0.07	0.58 ± 0.04	0.51 ± 0.06	0.56 ± 0.05
RF	0.72 ± 0.04	0.71 ± 0.06	0.66 ± 0.03	0.57 ± 0.05	0.66 ± 0.05
SVM	0.74 ± 0.05	0.73 ± 0.06	0.70 ± 0.04	0.67 ± 0.06	**0.71 ± 0.05**
XGB	0.64 ± 0.05	0.65 ± 0.06	0.63 ± 0.05	0.58 ± 0.06	0.62 ± 0.05

Table 7: Threshold $\tau = 0.6$. fairness results. Predictive accuracies across Race, Sex, and Migrant Status (MS) groups, highlighting mean disparities, standard deviations, and statistical significance.

	Model	AUC-PR	AUC-ROC	Accuracy	F1 Score
Race	LR	-0.09 ± 0.08 ***	-0.09 ± 0.06 ***	-0.15 ± 0.06 ***	-0.17 ± 0.08 ***
	MLP	-0.04 ± 0.08 **	-0.08 ± 0.07 ***	-0.07 ± 0.06 ***	-0.08 ± 0.08 ***
	RF	-0.10 ± 0.08 ***	-0.05 ± 0.06 ***	-0.14 ± 0.06 ***	-0.20 ± 0.08 ***
	SVM	-0.14 ± 0.08 ***	-0.11 ± 0.06 ***	-0.12 ± 0.06 ***	-0.13 ± 0.08 ***
	XGB	-0.12 ± 0.08 ***	-0.10 ± 0.06 ***	-0.11 ± 0.06 ***	-0.16 ± 0.08 ***
Sex	LR	-0.07 ± 0.08 ***	-0.10 ± 0.06 ***	-0.07 ± 0.06 ***	0.02 ± 0.08 *
	MLP	-0.09 ± 0.08 ***	-0.15 ± 0.07 ***	-0.14 ± 0.06 ***	-0.11 ± 0.08 ***
	RF	-0.07 ± 0.08 ***	-0.10 ± 0.06 ***	-0.12 ± 0.06 ***	-0.06 ± 0.08 ***
	SVM	-0.06 ± 0.07 ***	-0.09 ± 0.06 ***	-0.03 ± 0.06 ***	0.05 ± 0.07 ***
	XGB	-0.01 ± 0.08	-0.06 ± 0.06 ***	0.03 ± 0.06 ***	0.08 ± 0.08 ***
MS	LR	0.04 ± 0.08 **	-0.02 ± 0.06 **	-0.08 ± 0.06 ***	-0.04 ± 0.08 **
	MLP	-0.10 ± 0.10 ***	-0.11 ± 0.08 ***	-0.05 ± 0.06 ***	0.00 ± 0.09
	RF	0.00 ± 0.08	-0.04 ± 0.07 ***	-0.12 ± 0.06 ***	-0.10 ± 0.08 ***
	SVM	0.07 ± 0.08 ***	0.00 ± 0.07	-0.07 ± 0.06 ***	-0.04 ± 0.08 ***
	XGB	0.02 ± 0.08 *	-0.03 ± 0.08 **	-0.04 ± 0.07 ***	-0.05 ± 0.09 ***

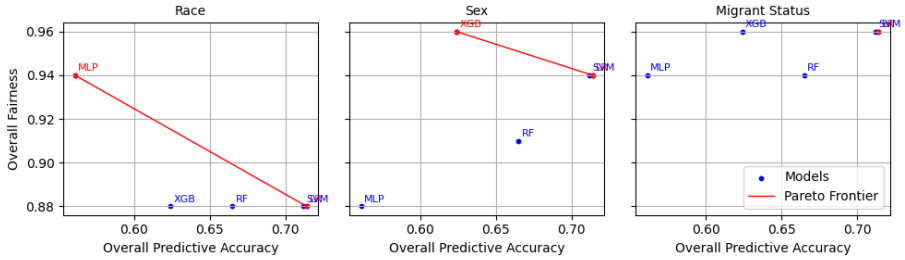

Fig. 2: Pareto frontier illustrating the trade-off between fairness and predictive accuracy. Red and blue points mark optimal and suboptimal models respectively for $\tau = 0.6$

Table 8: Threshold $\tau = 0.6$. RFG for various predictive accuracy metrics: PR(AUC-PR), ROC (AUC-ROC), Acc(Accuracy), and F1(F1 Score). Column initials represent demographic intersections: NWM (Non-White Migrant), WM (White Migrant), NWNM (Non-White Non-Migrant), WNM (White Non-Migrant), NWF (Non-White Female), NWM (Non-White Male), WF (White Female), WM (White Male), FM (Female Migrant), MM (Male Migrant), FNM (Female Non-Migrant), and MNM (Male Non-Migrant)

		Race-Migrant Status				Race-Sex				Sex-Migrant Status			
		NWM	WM	NWNM	WNM	NWF	NWM	WF	WM	FM	MM	FNM	MNM
LR	PR	-0.10	0.18	0.03	0.01	-0.03	0.02	0.05	0.07	0.07	0.08	0.01	0.03
	ROC	-0.08	0.12	0.00	0.01	-0.08	0.03	0.03	0.04	-0.07	0.05	-0.00	0.03
	Acc	-0.11	0.19	0.00	0.02	-0.05	-0.04	0.01	0.07	-0.07	-0.01	0.00	0.03
	F1	-0.11	0.20	-0.02	0.03	-0.05	-0.08	0.03	0.07	0.02	-0.03	0.01	0.01
MLP	PR	-0.13	0.16	0.05	0.01	-0.01	0.06	0.03	0.04	0.12	0.03	-0.02	0.07
	ROC	-0.18	0.03	0.04	0.00	-0.05	0.05	-0.03	0.05	-0.05	0.02	-0.03	0.06
	Acc	-0.04	0.07	-0.01	0.02	-0.07	0.05	0.01	0.03	-0.06	-0.01	-0.02	0.04
	F1	-0.20	0.19	0.01	-0.00	-0.05	0.03	0.01	0.05	0.06	-0.05	-0.04	0.04
RF	PR	-0.11	0.21	0.05	0.01	-0.06	0.01	0.04	0.04	0.05	0.06	-0.00	0.02
	ROC	-0.12	0.17	0.03	0.01	-0.10	0.05	0.03	0.00	-0.08	0.04	0.00	0.03
	Acc	-0.11	0.13	0.00	0.02	-0.09	0.02	0.02	0.04	-0.10	-0.04	-0.00	0.05
	F1	-0.15	0.18	-0.03	0.04	-0.16	0.03	0.07	0.02	-0.05	-0.05	0.01	0.04
SVM	PR	-0.05	0.15	0.01	0.03	-0.04	0.04	0.07	0.06	0.05	0.10	0.01	0.03
	ROC	-0.06	0.10	-0.01	0.02	-0.08	0.04	0.04	0.05	-0.08	0.09	-0.00	0.03
	Acc	-0.11	0.20	0.01	0.02	-0.04	-0.04	0.01	0.06	-0.07	-0.00	0.01	0.02
	F1	-0.12	0.21	-0.01	0.02	-0.03	-0.09	0.02	0.05	0.01	-0.02	0.02	-0.00
XGB	PR	-0.16	0.23	0.02	0.02	-0.00	-0.00	0.07	0.02	0.07	0.07	0.02	0.00
	ROC	-0.15	0.17	0.02	0.02	-0.05	0.01	0.02	0.02	-0.06	-0.00	0.00	0.03
	Acc	-0.11	0.26	0.01	0.01	-0.02	-0.04	0.02	0.03	-0.03	-0.01	0.02	-0.01
	F1	-0.21	0.30	0.00	0.02	-0.05	-0.08	0.04	0.04	0.03	-0.06	0.02	-0.01