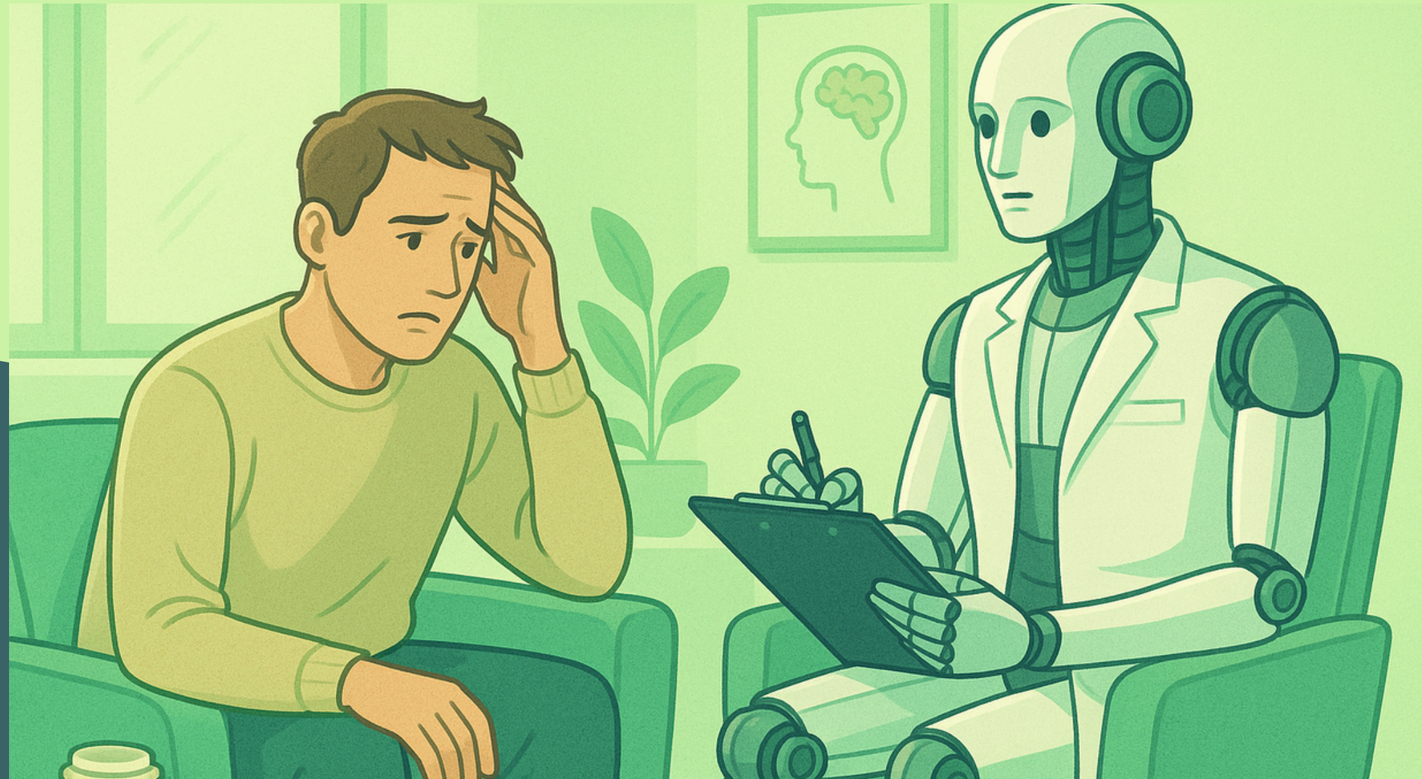
AUDIT OF DEPRESSION PREDICTION ADS

by Era Sarda & Judy Yang
es6790@nyu.edu | hy1331@nyu.edu



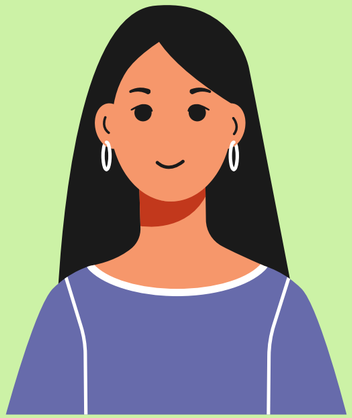
Content

- Background and Motivation
- Stakeholders
- About the Data
- Model Architecture
- Fairness Audit
- ThresholdOptimizer
- LIME

Background and Motivation

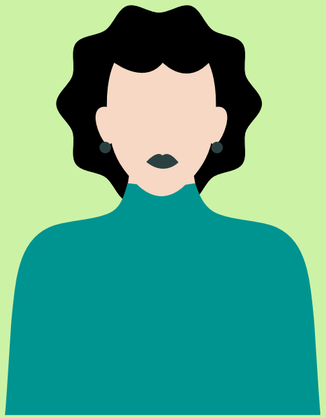
- Depression prediction model
  - depression shows gender disparities in reported prevalences
  - but the gap is smaller or absent in developing countries
  - the difference indicates possible bias in how data is recorded and/or interpreted
- ADS in the context of mental health prediction
  - models trained on biased data can replicate or amplify social inequalities
  - misclassifications have real-life consequences
- Kaggle Competition - model by Mahdi Ravaghi:
    - won Kaggle’s Playground Series S4E11
    - highest accuracy among 2,500+ submissions
    - evaluated solely on overall accuracy
- Our Goal
  - audit this high-performing model for fairness
  - does it treat subgroups equitably?
  - can we mitigate potential bias without compromising model performance?

Stakeholders



Patient
&
Family

Clinicians



Employers
&
Educational
Institutes

Researchers

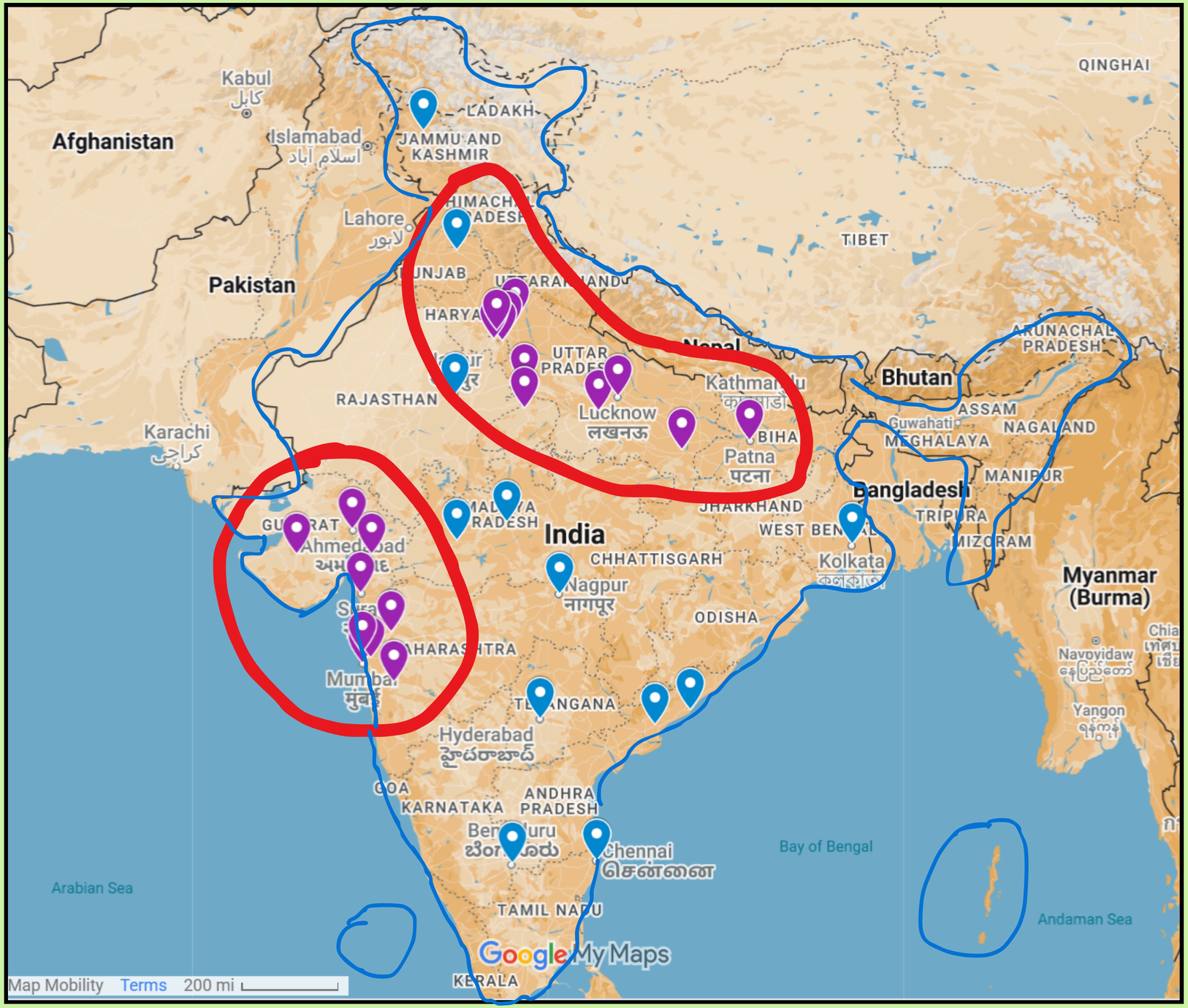
Mental Health
Advocacy
Organizations

Government

ATTRIBUTES

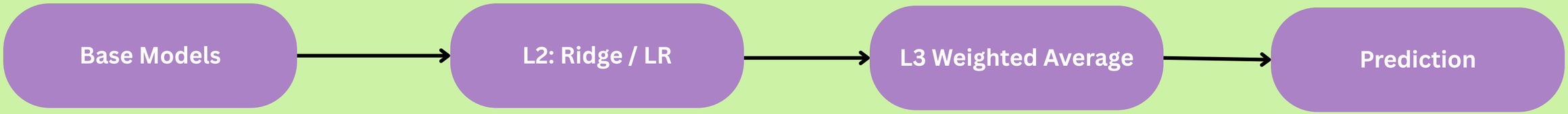
- Name
- Age
- Gender
- City
- Degree
- Dietary Habits
- Sleep Duration
- Suicidal Thoughts
- Working Professional or Student
- Work, Academic Pressure
- Job, Study Satisfaction
- Profession
- CGPA

Regional Distribution of Data



The Model

- an ensemble classification model built to predict depression and maximize accuracy
- Model Architecture:
  - Base models:
    - AdaBoost
    - CatBoost
    - XGBoost
    - Gradient Boost
    - LGBMClassifier
    - Logistic Regression
  - L2 ensemble layer:
    - Logistic Regression
    - Ridge Regression
  - L3 ensemble: weighted average of L2 outputs optimized for accuracy



ATTRIBUTES

~~Name~~

Age

Gender

~~City~~

Degree

Dietary Habits

Sleep Duration

Suicidal Thoughts

Working Professional or Student

Work, Academic Pressure

Job, Study Satisfaction

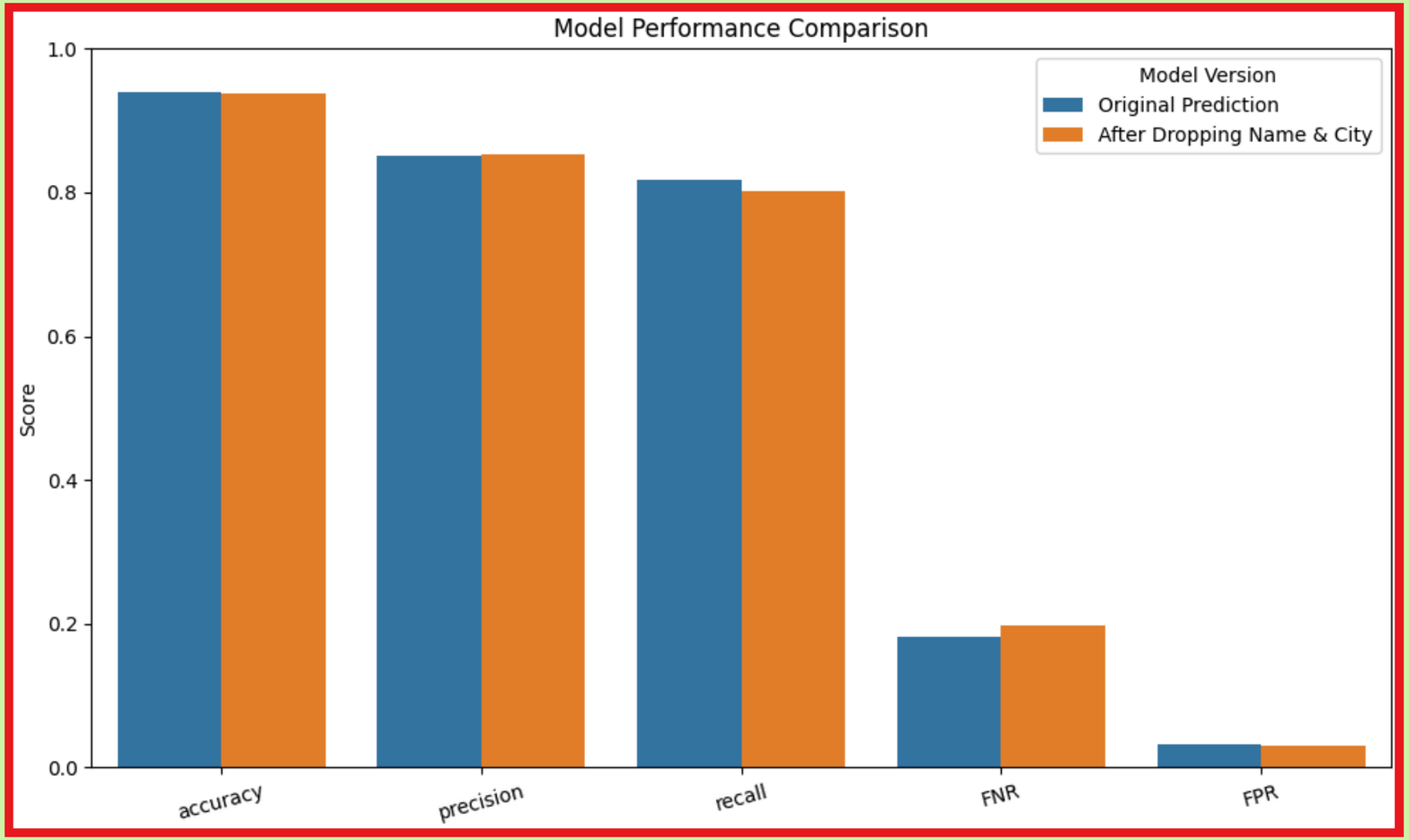
Profession

CGPA

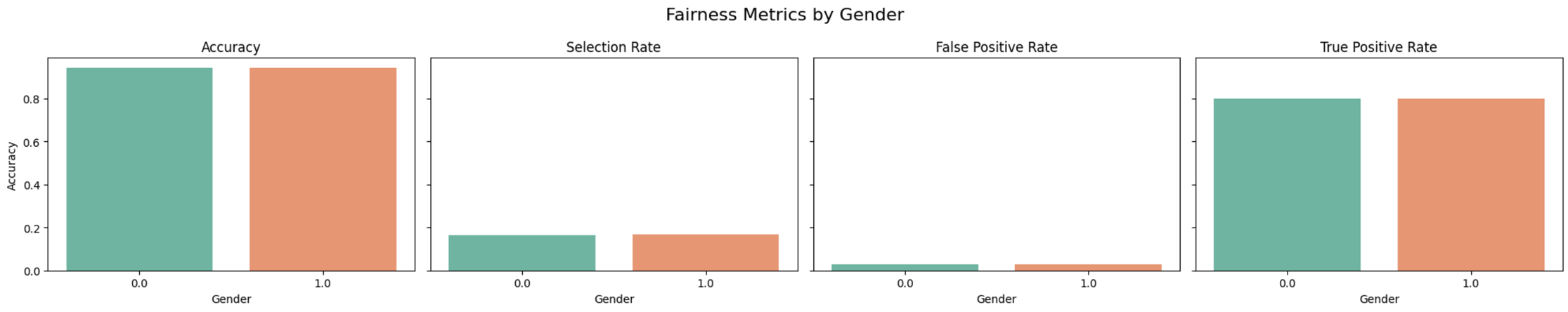
Removing Sensitive Attributes

Name

City



Fairness Metric by Gender



- Model performance is nearly **identical** between genders:
  - Accuracy, TPR, FPR, and Selection Rate are all balanced
  - no meaningful evidence of gender disparity in the model
- Despite well-documented gender differences in depression prevalence, this model does not appear biased by gender
- This result prompted us to explore other features as sensitive features

ATTRIBUTES

~~Name~~

Age

~~Gender~~

~~City~~

Degree

Dietary Habits

Sleep Duration

Suicidal Thoughts

Working Professional or Student

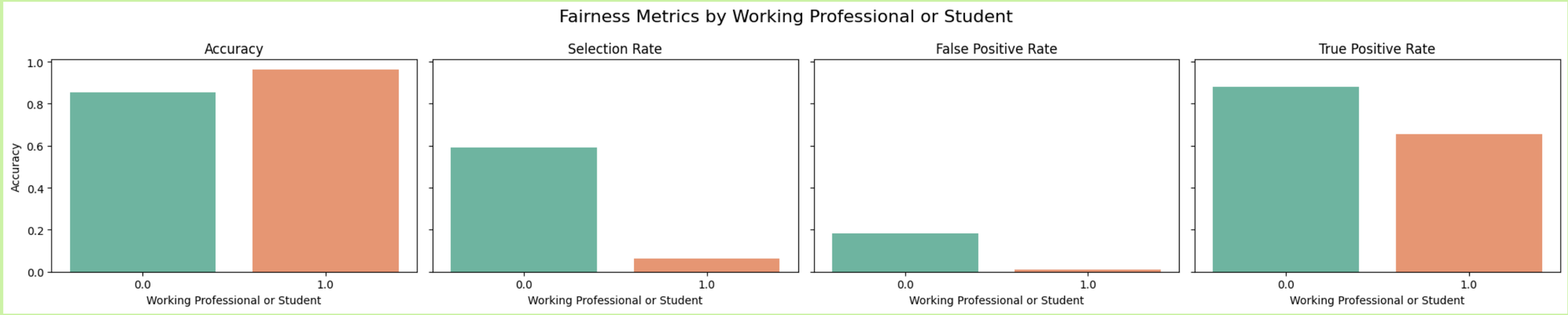
Work, Academic Pressure

Job, Study Satisfaction

Profession

CGPA

Fairness Metric by Employment Status



- EDA revealed strong association between age, employment status (student or working professional) and depression rate
- Student (group 0):
  - Much higher Selection Rate
  - Higher FPR and TPR
- Working Professional (group 1):
  - Higher overall accuracy
  - Low Selection and FPR
  - **Low TPR**: true positive often missed
- Model is over-identifying depression in student and under-identifying in working professionals - employment status is a meaningful sensitive feature

ATTRIBUTES

- Name
- Age
- Gender
- City
- Degree
- Dietary Habits
- Sleep Duration
- Suicidal Thoughts

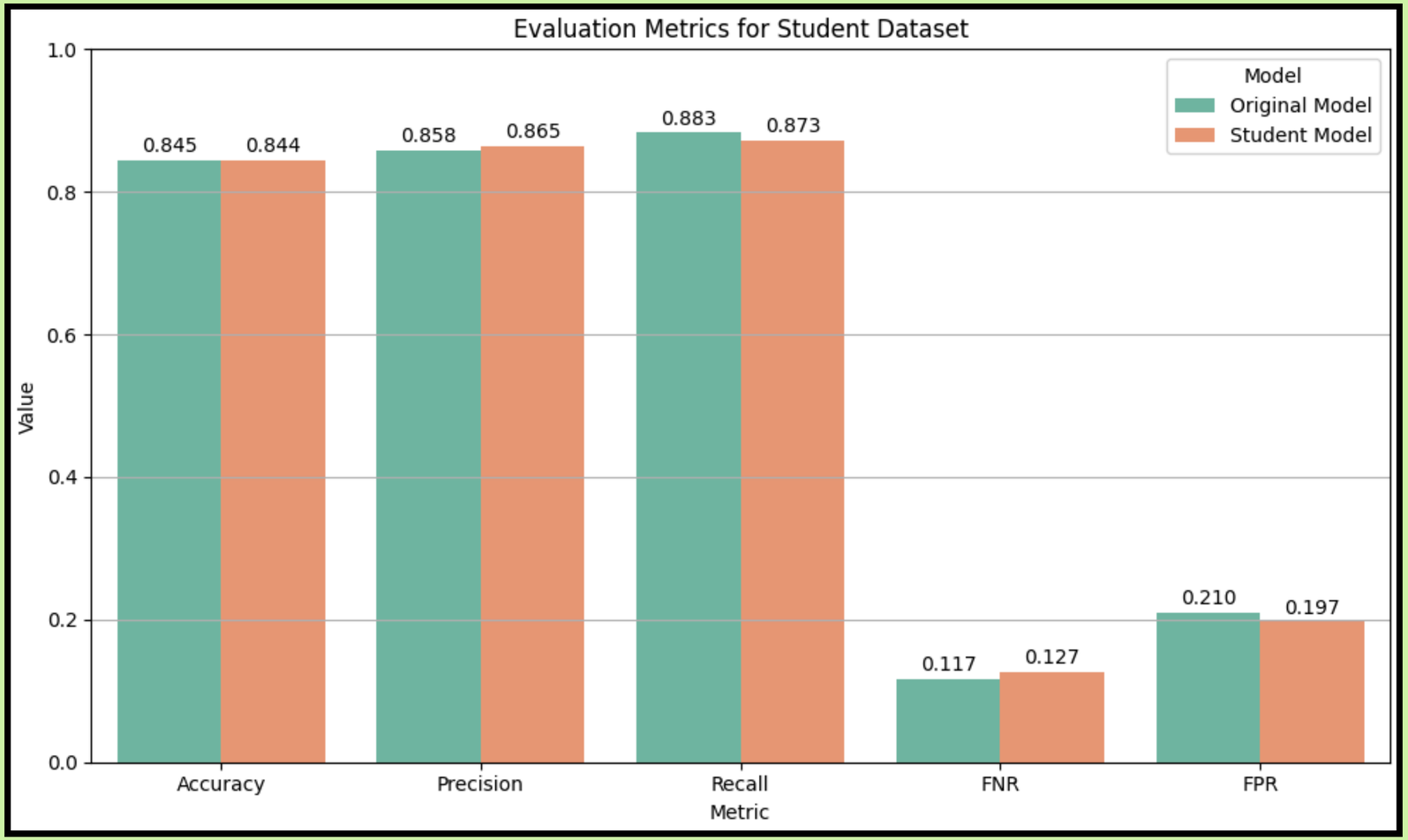
~~Working Professional or Student~~

~~Work~~, Academic Pressure

~~Job~~, Study Satisfaction

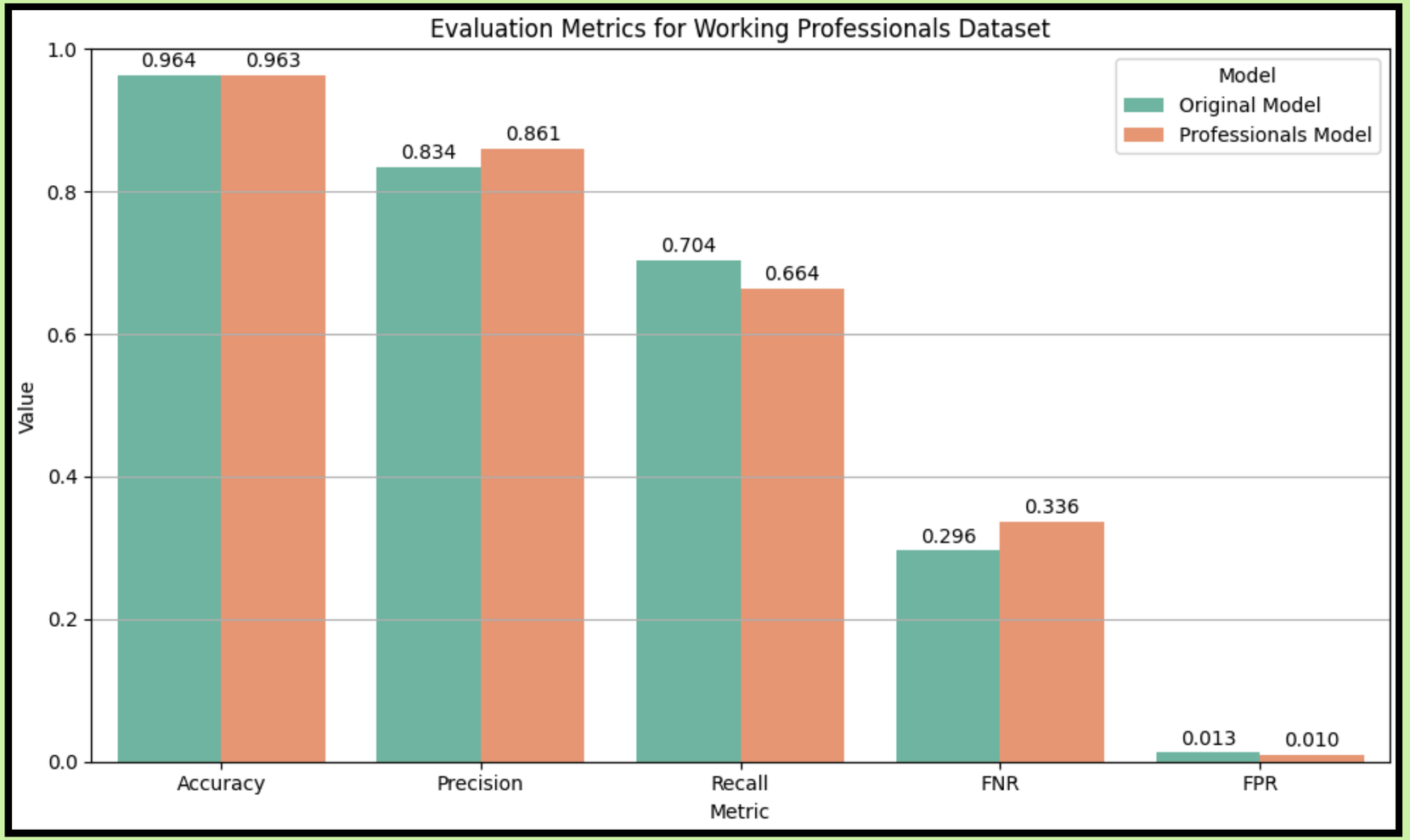
~~Profession~~

CGPA


Evaluation Metrics for Student Dataset
Model
Original Model
Student Model
Value
Metric
1.0
0.8
0.6
0.4
0.2
0.0
0.845
0.844
0.858
0.865
0.883
0.873
0.117
0.127
0.210
0.197
Accuracy
Precision
Recall
FNR
FPR

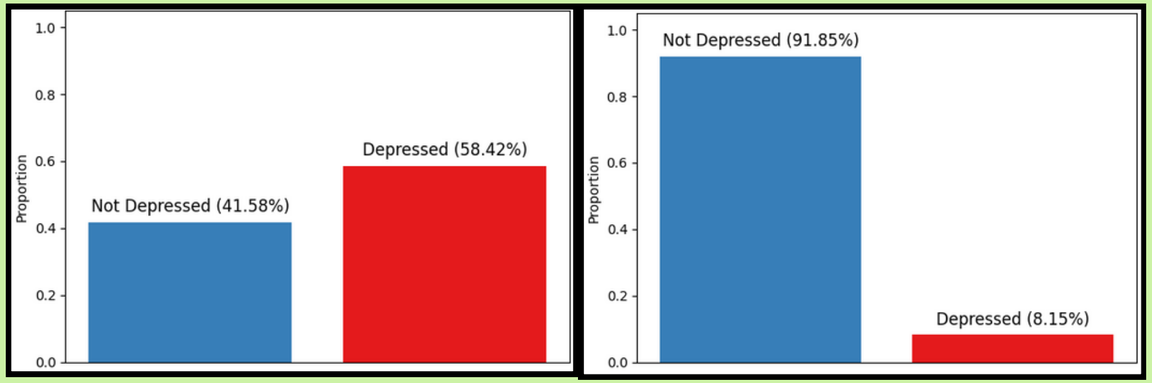
ATTRIBUTES

- Name
- Age
- Gender
- City
- Degree
- Dietary Habits
- Sleep Duration
- Suicidal Thoughts
- ~~Working Professional or Student~~
- Work, ~~Academic~~ Pressure
- Job, ~~Study~~ Satisfaction
- Profession
- ~~CGPA~~


Evaluation Metrics for Working Professionals Dataset
Model
Original Model
Professionals Model
Value
Metric
1.0
0.8
0.6
0.4
0.2
0.0
0.964
0.963
0.834
0.861
0.704
0.664
0.296
0.336
0.013
0.010
Accuracy
Precision
Recall
FNR
FPR

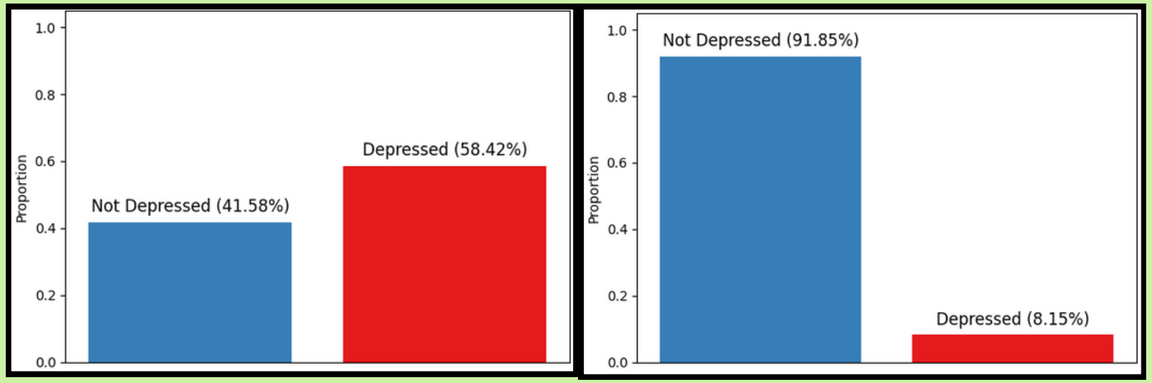
Then what’s causing the disparity in FNR and FPR in both the groups?

Ground Truth Label Disparity?



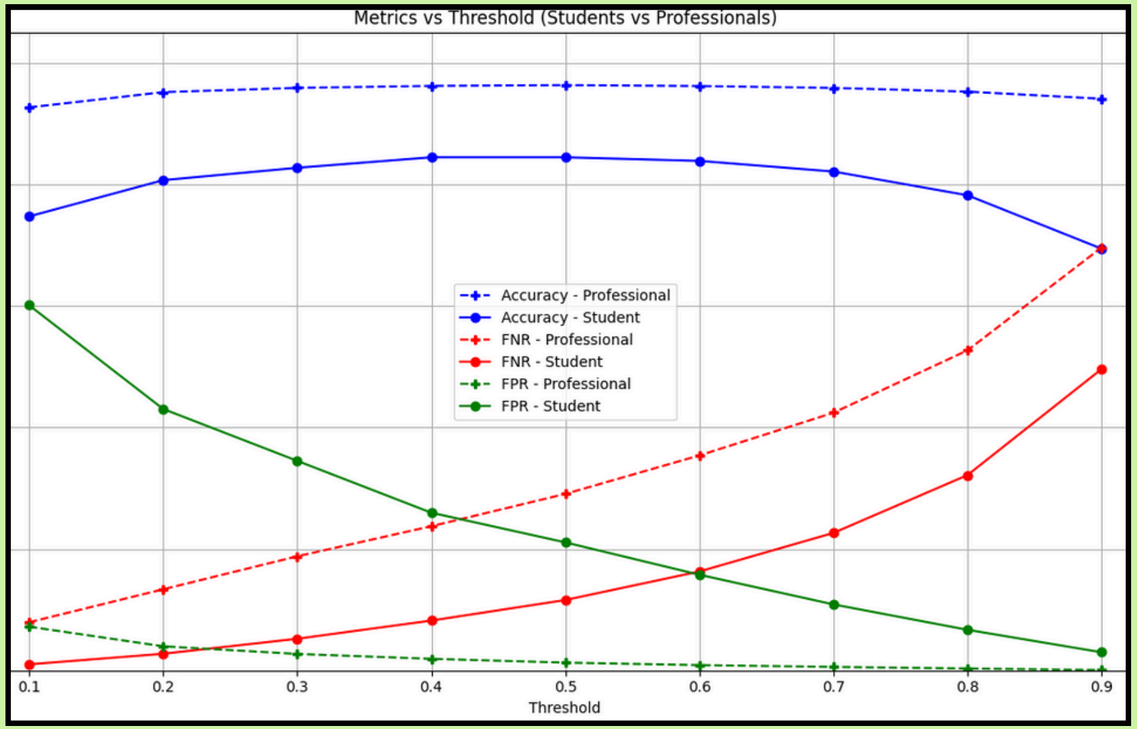
Then what's causing the disparity in FNR and FPR in both the groups?

Ground Truth Label Disparity?



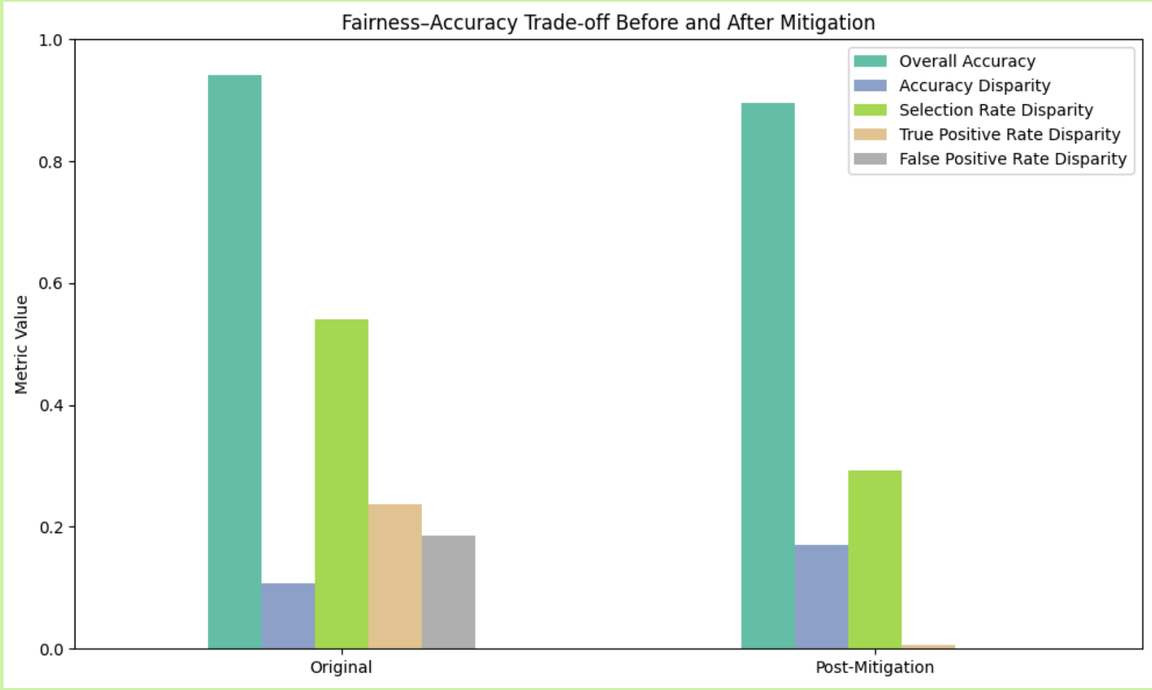
Then what's causing the disparity in FNR and FPR in both the groups?

Threshold for the Binary Classification?



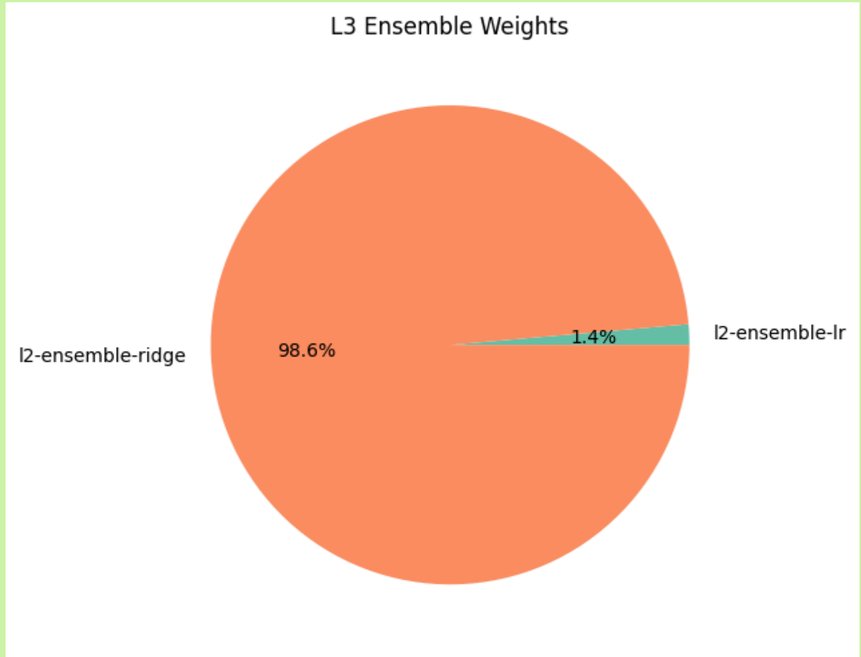
Mitigating Disparity: ThresholdOptimizer

- Fairlearn’s ThresholdOptimizer is a post-processing fairness mitigation tool - **adjust decision threshold** and enforce fairness constraints
- Goal: equalize true and false positive rates across groups without restraining the model ⟶ equalized odds
- After mitigation:
  - TPR is more aligned
  - FPR converged
  - Selection Rate disparity decreased
  - Slight drop in overall accuracy



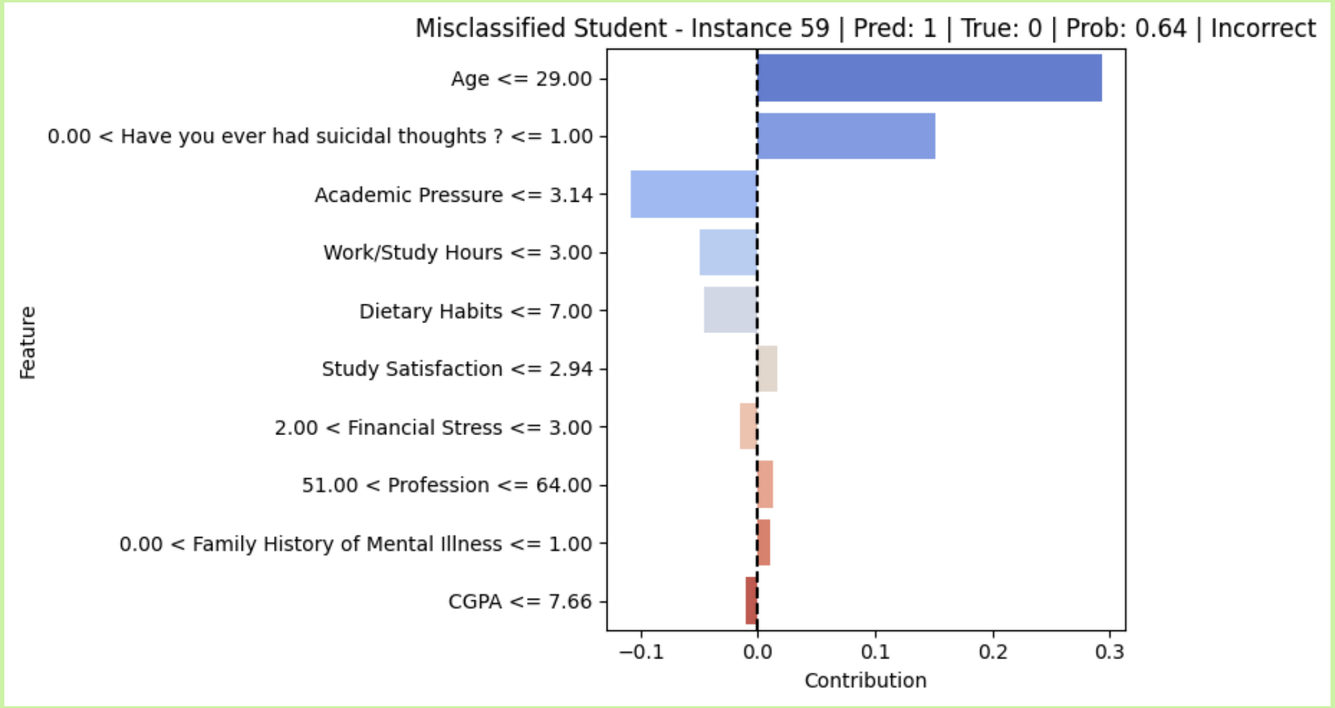
Local Interpretable Model-Agnostic Explanations (LIME)

- Explains why the model made a specific prediction
- Why LIME:
  - our model is an ensemble, which makes it hard to interpret globally
  - which features are the most important in prediction
  - human-interpretable explanations
- We used LIME to explain predictions from the L2 LR component of the ensemble
- Three instances:
  - Misclassified student
  - Misclassified working professional
  - Uncertain prediction



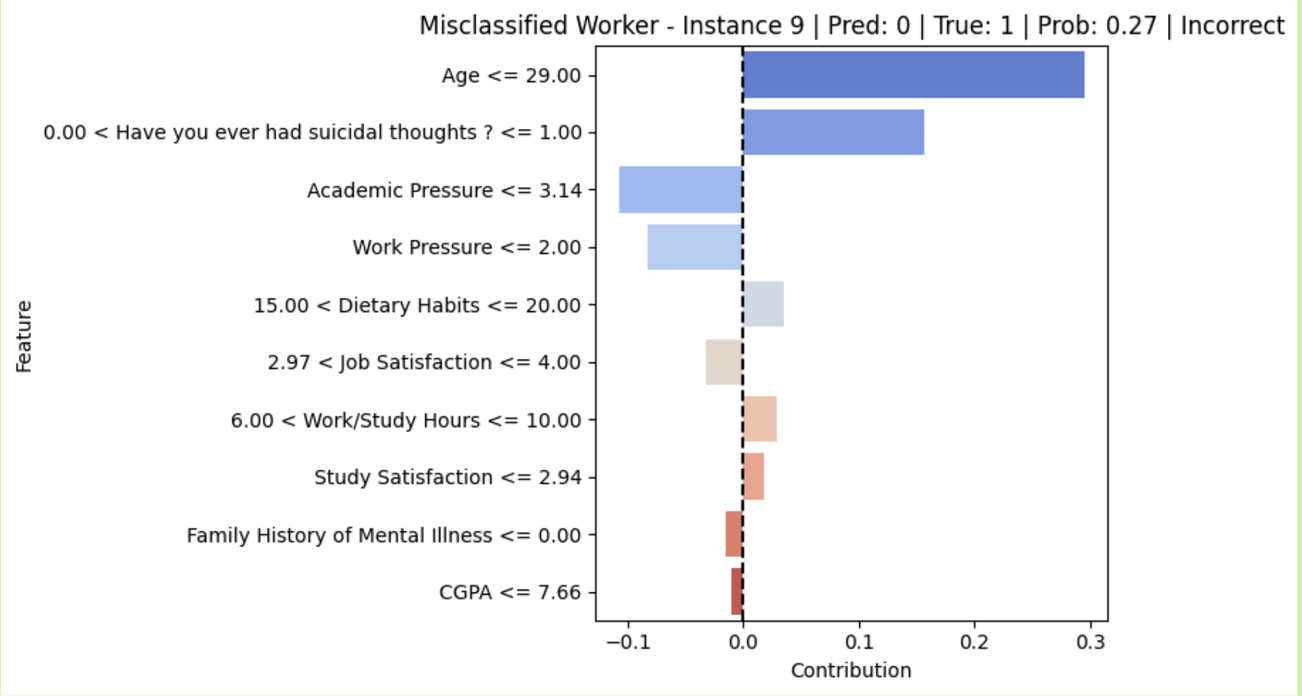
LIME: Misclassified Student

- True label: not depressed
- Predicted: depressed
- Strong positive contributions from:
  - young age
  - reported suicidal thought
  - high academic pressure
- reflects a pattern of over-identification in students
- Model relies on generic high-risk factors (age)
- Might be beneficial to err on the side of caution, but might incur unnecessary interventions or stigma for the student



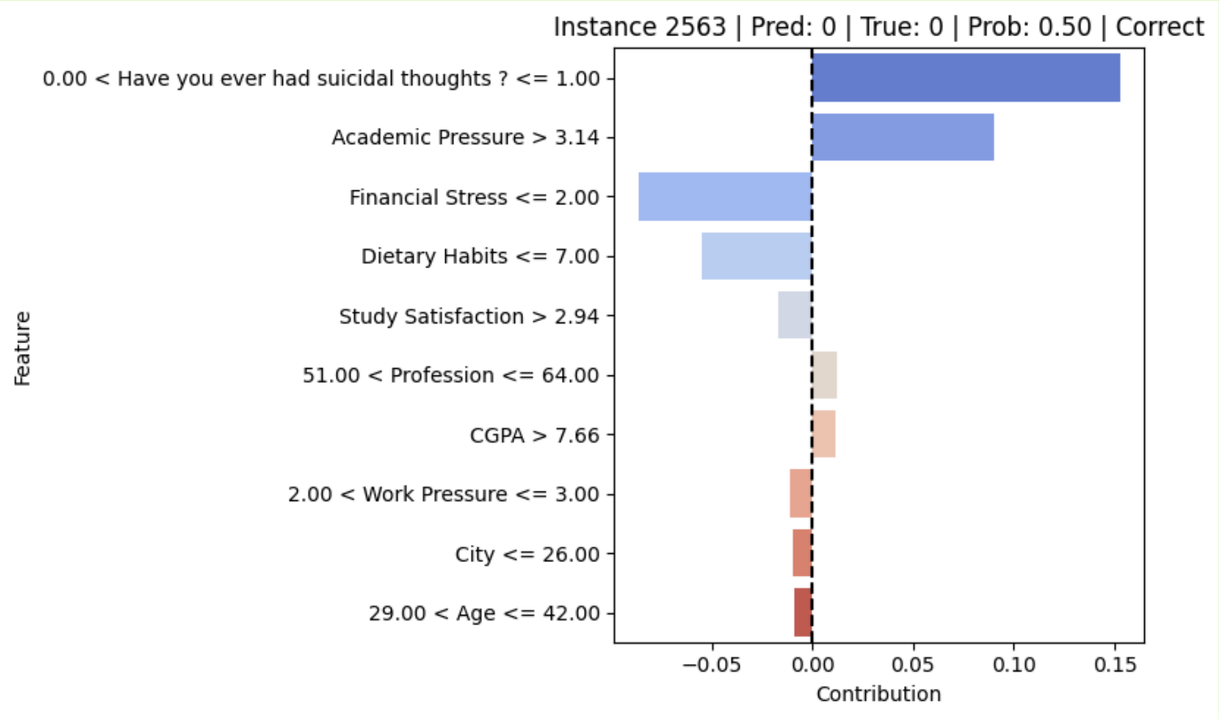
LIME: Misclassified Working Professional

- True label: depressed
- Predicted: not depressed
- Strong positive contributions from:
  - young age
  - suicidal thought
- reflects a pattern of under-identification in working professionals
- possible bias toward ignoring emotional distress in older / working individuals
- Could have serious consequences for the individual and their community



LIME: Uncertain Instance

- True label: depressed
- Predicted: depressed
- Small positive contributions from:
  - suicidal thoughts
  - moderate academic pressure
- No feature contributed strongly in either direction
- Model shows nuanced behavior when signals are ambiguous
- Potential fragility near decision boundary



Conclusion and Takeaways

- Findings
  - Minimal gender bias
  - Major disparities in employment status:
    - Selection Rate
    - TPR and FPR
    - model over-identifies students and misses working professionals
- Method
  - Fairlearn’s MetricFrame and ThresholdOptimizer to measure and mitigate bias
  - LIME to interpret predictions at the individual level
- Takeaways
  - high accuracy ≠ fair model
  - stakeholders: clinicians, policy makers, patients and their family/community
  - data collection for such ADS must ensure they reflect practical scenarios
  - Depression is a sensitive topic that requires not only accurate but also equitable and explainable predictions

Thank You!