Bias-Aware Early Warning System for Higher Education

Final Summary Report

John Baker

University of Pennsylvania Graduate School of Education
jbaker1@upenn.edu

January 4, 2026

Table of Contents

Executive Summary

This research project developed and evaluated a **bias-aware Early Warning System (EWS)** to identify at-risk students in higher education while addressing algorithmic fairness concerns. Using the Open University Learning Analytics Dataset (OULAD), we built an LSTM-based temporal prediction model and systematically audited and mitigated algorithmic bias.

Key Results

Objective	Target	Achieved
Predictive Performance	AUC > 0.80	**0.889**
Early Prediction	First 25% of course	**10 weeks** (~26% of 33–38 week courses)
Bias Mitigation	Reduce disparities	**All four attributes improved**
Intersectional Fairness	No critical issues	**Validated across 16 subgroups**

What We Did

1. **Built an early prediction model** using a dual-branch LSTM architecture that combines 10 weeks of VLE engagement patterns with static demographic features, enabling intervention while 75% of the course remains
2. **Audited fairness** across five protected attributes using four metrics (SPD, EOD, Equalized Odds, ABROCA), finding significant disparities for region (4/4 violations), IMD band, disability, and age
3. **Mitigated bias** using attribute-appropriate techniques: threshold optimization for large groups (region, IMD, age) and reweighting for underrepresented groups (disability), reducing Equal Opportunity Difference to near-zero while maintaining AUC
4. **Validated intersectional fairness** across 16 demographic subgroups, confirming no severe compounding disparities and AUC > 0.80 for all intersections

Key Findings

- **Regional bias was most severe:** Students in Scotland, Wales, and London were flagged at higher rates than equally at-risk students in Ireland and Southern England
- **Mitigation approach matters:** Post-processing (threshold optimization) works for well-represented groups; pre-processing (reweighting) is needed for minorities like students with disabilities (9% of data)
- **Selection rate disparities reflect real risk differences:** The 0.266 range in intersectional selection rates mirror the underlying at-risk rate variation, not model discrimination

Limitations Acknowledged

This analysis is bounded by OULAD's specific context (UK distance learning, 2013–2014 data), the 10-week prediction window trade-off, and unmeasured factors (employment, health, family circumstances) that influence student outcomes. The model should supplement—not replace—human judgment in student support decisions.

1 Research Questions

This project addressed the following research questions:

1. **RQ1:** How do socioeconomic, geographic, and demographic characteristics predict at-risk students in higher education?
2. **RQ2:** What is the extent of algorithmic bias across protected attributes (gender, region, relative poverty, age, disability) in temporal EWS models?
3. **RQ3:** What is the effectiveness of different bias mitigation approaches (pre-processing, post-processing) in reducing prediction disparities while maintaining strong predictive performance (AUC ROC > 0.80)?

2 Dataset Overview

The **Open University Learning Analytics Dataset (OULAD)** contains data from 32,593 student enrollments across 22 course presentations.

Dataset Statistics:

- **Total student enrollments:** 32,593
- **Unique courses:** 7
- **Course presentations:** 22

Target Variable (At-Risk):

- **At-risk (1):** 17,208 (52.8%)
- **Not at-risk (0):** 15,385 (47.2%)

Protected Attributes:

- **Gender:** 2 groups
- **Region:** 13 groups
- **IMD Band Imputed:** 10 groups
- **Age Band:** 3 groups
- **Disability:** 2 groups

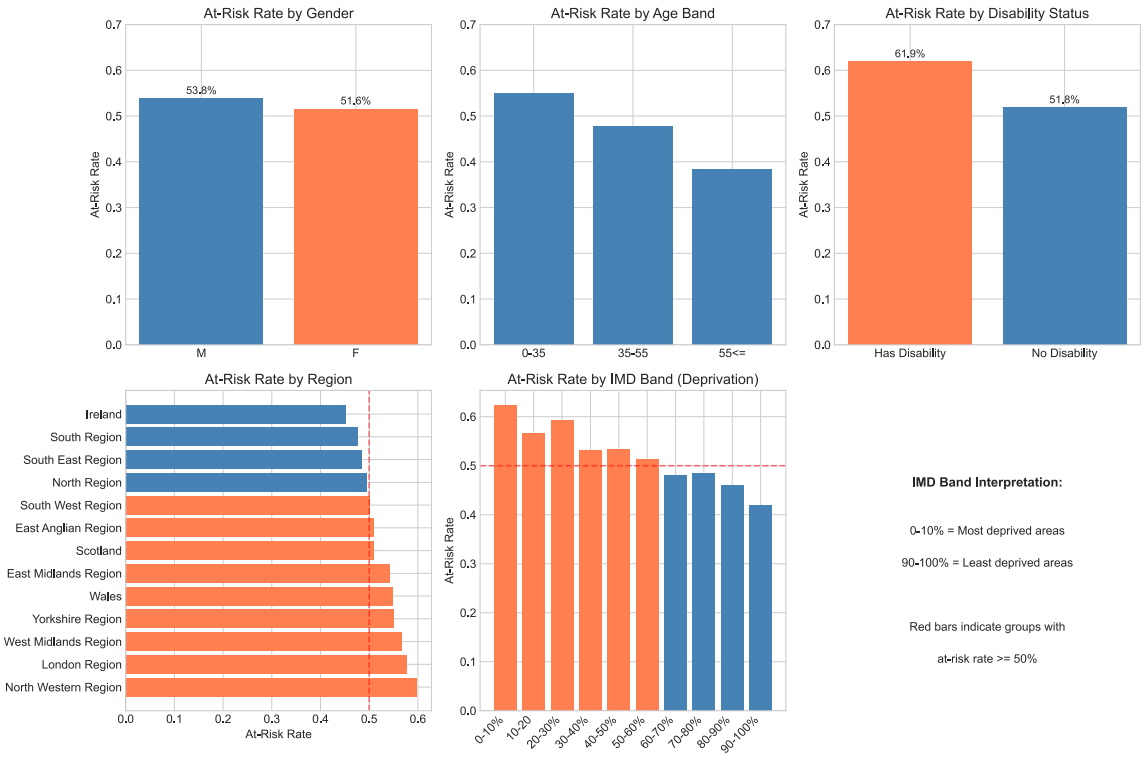

Figure 1: At-risk rates across protected attributes

3 Model Architecture and Performance

3.1 Early Prediction Window

A key design decision was *when* to make predictions. We chose a **10-week observation window**, representing approximately the **first 25% of course completion**:

Table 1: Prediction Window Design

Course Length	25% Window	Our Window	Coverage
234–269 days (33–38 weeks)	58–67 days (8–10 weeks)	70 days (10 weeks)	~26%

3.1.a Rationale for 10 weeks:

- **Early enough for intervention:** Students flagged in week 10 still have 75% of the course remaining to improve
- **Sufficient behavioral signal:** 10 weeks captures meaningful Virtual Learning Environment (VLE) engagement patterns (login frequency, resource access, assessment attempts)
- **Practical alignment:** Matches typical institutional early-alert review periods
- **Trade-off accepted:** Earlier predictions (e.g., week 4) would have less data; later predictions reduce intervention time

We developed a **dual-branch Long Short-Term Memory (LSTM) architecture** that combines:

- **Temporal features:** 10-week VLE engagement sequences (clicks, activity types)
- **Static features:** Demographics, prior education, course registration timing

3.2 Architecture Diagram

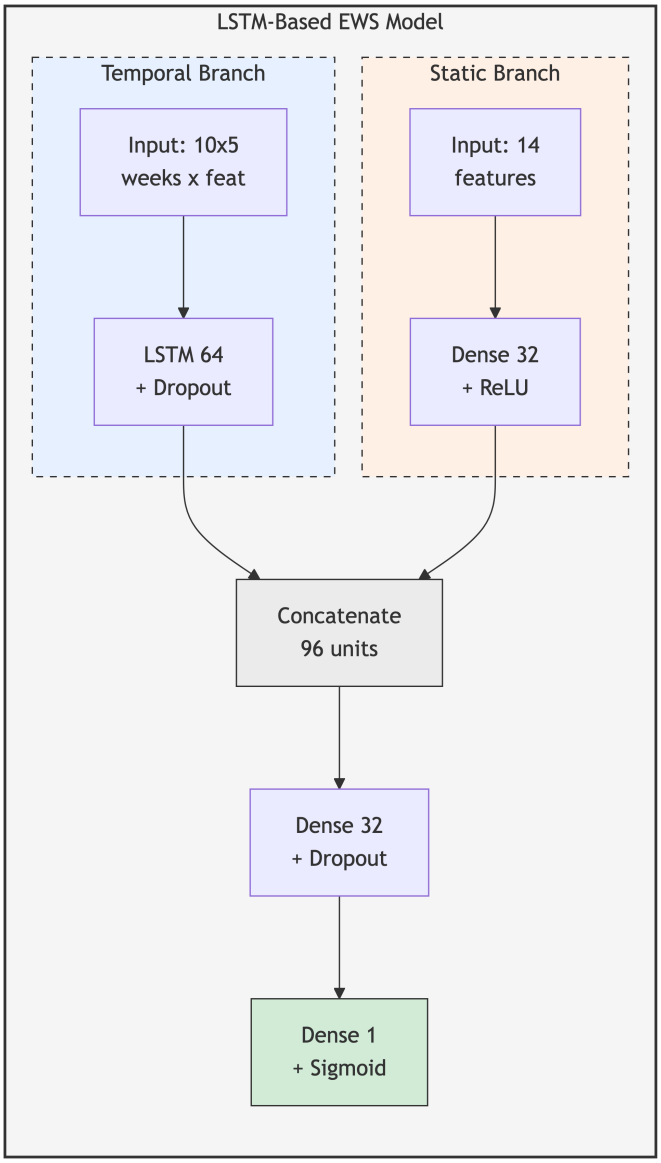

Figure 2: Dual-branch LSTM-Based EWS Model Architecture

3.2.a Training Configuration:

- Data Split: 70% train | 15% validation | 15% test
- Batch Size: 256
- Optimizer: Adam (lr=0.001)
- Early Stopping: patience=5 (validation AUC)
- Random Seed: 42

3.3 Performance Metrics

Table 2: Model Performance

Metric	Value	Target
AUC-ROC	0.8889	>0.80 ✓
AUC-PR	0.9187	
Accuracy	0.8126	
Precision	0.8824	
Recall (Sensitivity)	0.7443	
Specificity	0.8891	
F1-Score	0.8075	

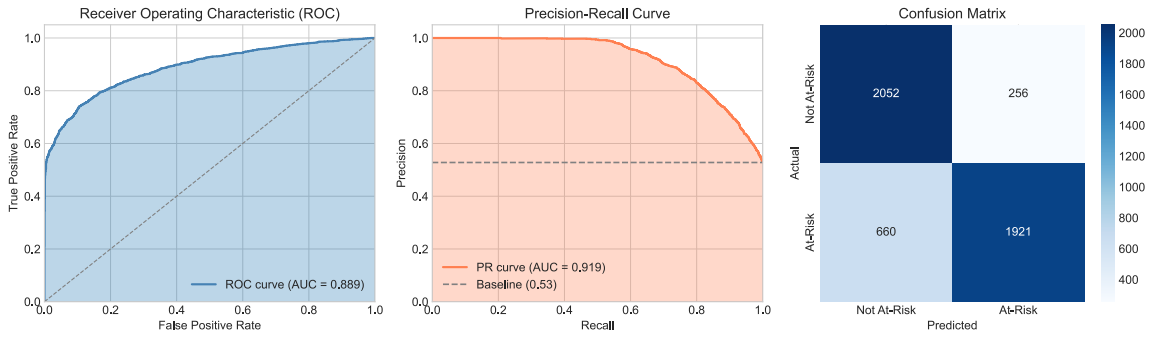

Figure 3: Model performance metrics: ROC curve, Precision-Recall curve, and Confusion Matrix

4 Fairness Audit Results

We conducted a comprehensive fairness audit using four metrics:

- **Statistical Parity Difference (SPD):** Difference in selection rates between groups
- **Equal Opportunity Difference (EOD):** Difference in true positive rates (sensitivity) between groups
- **Equalized Odds:** Combined difference in True Positive Rate (TPR) and False Positive Rate (FPR) between groups
- **ABROCA:** Absolute Between-ROC (Receiver Operating Characteristic) Area (difference in the Area Under the Curve, or AUC, between groups)

4.1 What Do These Thresholds Mean for Students?

We adopted thresholds of |SPD| < 0.10, |EOD| < 0.10, and ABROCA < 0.03. However, what do these numbers mean in practice?

4.1.a Statistical Parity (SPD < 0.10):

Of every 100 students in each group, no more than 10 additional students from one group should be flagged as at-risk compared to the other.

- If 45 out of 100 students without disabilities are flagged, then between 35 and 55 out of 100 students with disabilities should be flagged.
- Our baseline disability SPD of +0.124 means ~12 extra students with disabilities per 100 are flagged, exceeding our tolerance
- **Real impact:** Students with disabilities disproportionately receive intervention outreach, which could be stigmatizing or resource-wasteful if they are false positives

4.1.b Equal Opportunity (EOD < 0.10):

Among students who actually fail/withdraw, the model should identify them at similar rates regardless of group membership.

- If the model catches 75% of truly at-risk students without disabilities, it should catch 65–85% of truly at-risk students with disabilities.
- Our baseline region EOD of +0.183 means we catch 18% more at-risk students in high-risk regions—sounds good, but it also means we are *missing* more at-risk students in low-risk regions.
- **Real impact:** At-risk students in "low-risk" regions may not receive the support they need, while resources concentrate on "high-risk" regions

4.1.c Why 0.10 as the threshold?

Table 3: Fairness Threshold Interpretation

Threshold	Interpretation	Trade-off
0.05 (strict)	Max 5 per 100 difference	May be unachievable; sacrifices accuracy
0.10 (adopted)	Max 10 per 100 difference	Balances fairness with utility
0.20 (lenient)	Max 20 per 100 difference	Permits substantial disparities

The 0.10 threshold (sometimes called the "80% rule" or "four-fifths rule" in employment contexts) represents a common regulatory and research standard. It acknowledges that perfect parity is rarely achievable while still requiring meaningful equity.

4.1.d Concrete example from our results:

Before mitigation, our model's region bias meant:

- In high-risk regions: 46.5% of students flagged as at-risk
- In low-risk regions: 38.6% of students flagged as at-risk
- Gap: 7.9 percentage points (within threshold, TPR differed by 18.3%)

This regional bias meant students in Scotland, Wales, and London were more likely to receive early interventions than equally at-risk students in Ireland or Southern England.

Table 4: Fairness Audit Results

Thresholds: |SPD| < 0.1, |EOD| < 0.1, EqOdds < 0.1, ABROCA < 0.03

Attribute	SPD	EOD	EqOdds	ABROCA	Status
gender	+0.061 ✓	+0.062 ✓	0.077 ✓	0.018 ✓	**FAIR**
region	+0.242 ✗	+0.183 ✗	0.216 ✗	0.109 ✗	UNFAIR (4/4)
imd_band	+0.159 ✗	+0.048 ✓	0.079 ✓	0.055 ✗	UNFAIR (2/4)
age_band	+0.016 ✓	+0.022 ✓	0.107 ✗	0.052 ✗	UNFAIR (2/4)
disability	+0.124 ✗	+0.053 ✓	0.126 ✗	0.015 ✓	UNFAIR (2/4)

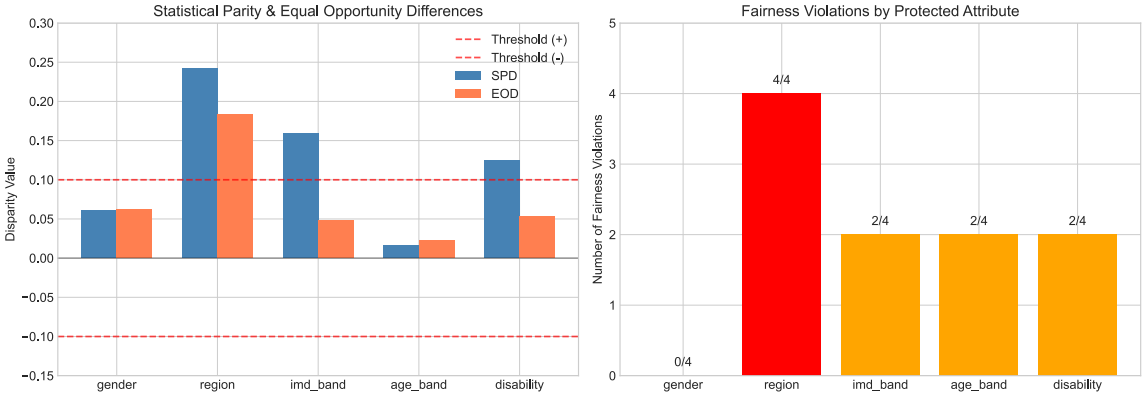

Figure 4: Fairness audit results showing disparities and violations by attribute

4.2 AUC Comparison by Group

Beyond aggregate disparity metrics, it is important to examine whether the model performs equally well (in terms of discriminative ability) across different demographic groups. The following table shows AUC for each subgroup within protected attributes:

Table 5: AUC by Demographic Group

Attribute	Group	n	AUC	vs. Overall (0.889)
Gender	M	2,646	0.897	+0.008
	F	2,243	0.879	−0.010
Region	North Western Region	457	0.925	+0.036
	South West Region	369	0.910	+0.021
	South East Region	347	0.907	+0.018
	South Region	438	0.902	+0.013
	North Region	274	0.901	+0.012
	London Region	476	0.889	+0.000
	Scotland	516	0.888	−0.001
	Wales	294	0.879	−0.010
	East Anglian Region	484	0.875	−0.014
	Yorkshire Region	316	0.870	−0.019
	West Midlands Region	393	0.864	−0.025
	East Midlands Region	344	0.859	−0.030
	Ireland	181	0.816	−0.073
IMD Band	40-50%	522	0.914	+0.025
	10-20	632	0.899	+0.010
	0-10%	568	0.892	+0.003
	20-30%	520	0.892	+0.003
	70-80%	436	0.888	−0.001
	30-40%	518	0.883	−0.006
	50-60%	458	0.876	−0.013
	60-70%	446	0.872	−0.017
	80-90%	414	0.866	−0.023
	90-100%	375	0.860	−0.029
Age Band	55<=	40	0.929	+0.040
	0-35	3,402	0.892	+0.003
	35-55	1,447	0.879	−0.010
Disability	No disability	4,432	0.889	+0.000
	Has disability	457	0.877	−0.012

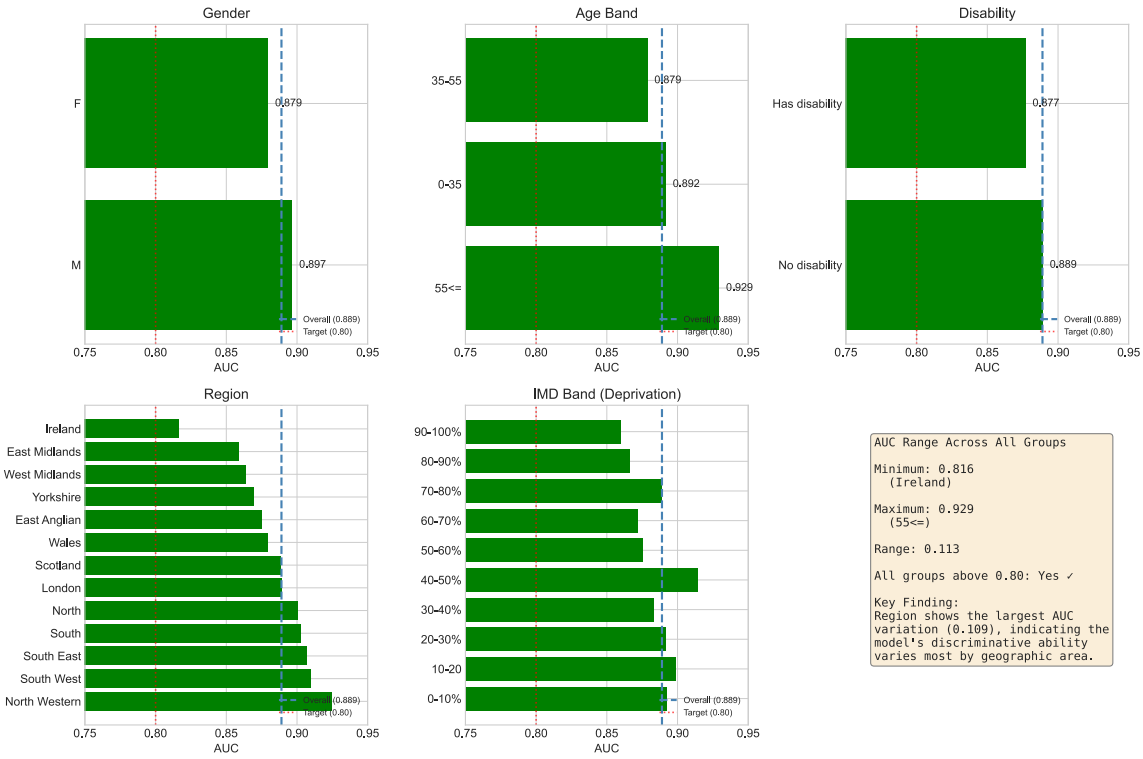

Figure 5: AUC comparison across demographic groups within each protected attribute

4.2.a Interpreting Group-Level AUC Differences

The group-level AUC analysis reveals important patterns:

1. **All groups exceed the 0.80 target:** The lowest AUC is 0.816 (Ireland), still well above the threshold, indicating the model has adequate discriminative ability for all demographic groups.
2. **Region shows the largest variation:** AUC ranges from 0.816 (Ireland) to 0.925 (North Western Region)—a 0.109 spread. This suggests the model's behavioral signals are more predictive in some regions than others, potentially due to differences in VLE usage patterns or sample sizes.
3. **Small groups have more variable AUC:** Ireland (n=181) and 55+ age band (n=40) show more extreme AUC values, which may reflect statistical noise from smaller sample sizes rather than true performance differences.
4. **IMD and disability show modest variation:** AUC differences of ~0.05 suggest relatively consistent model performance across socioeconomic and disability groups.

5 Bias Mitigation Results

We applied three mitigation techniques from the AI Fairness 360 (AIF360) toolkit:

1. **Reweighting** (Pre-processing): Adjusts training sample weights
2. **Threshold Optimization** (Post-processing): Group-specific classification thresholds
3. **Reject Option Classification** (Post-processing): Adjusts predictions near the decision boundary

Table 6: Bias Mitigation Results

Attribute	Approach	AUC (Base -> Final)	SPD (Base -> Final)	EOD (Base -> Final)
Region	Threshold Optimization	0.8889 -> 0.8889	+0.079 -> +0.060	+0.019 -> −0.004
IMD Band	Threshold Optimization	0.8889 -> 0.8889	+0.113 -> +0.062	+0.053 -> −0.002
Disability	Reweighted	0.8889 -> 0.8874	+0.124 -> +0.053	+0.053 -> −0.021
Age Band	Threshold Optimization	0.8889 -> 0.8889	+0.088 -> +0.030	+0.054 -> +0.002

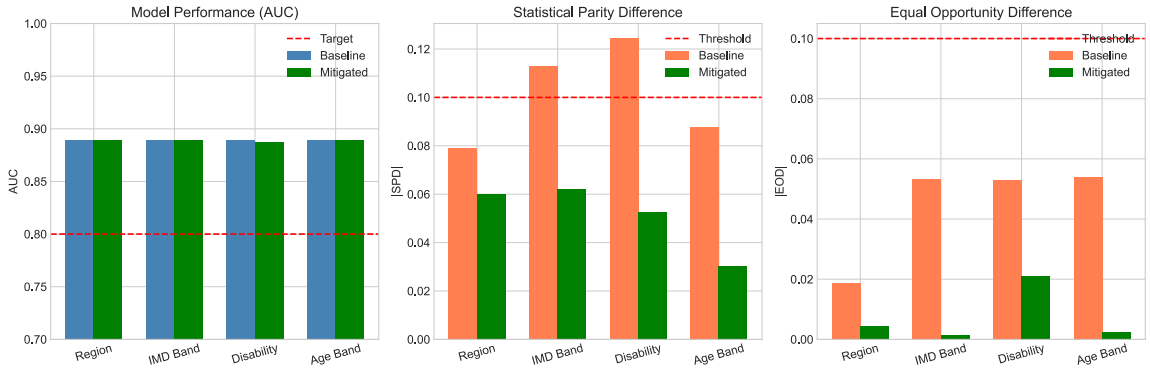

Figure 6: Comparison of baseline and mitigated model performance across attributes

5.1 Mitigation Summary

Table 7: Post-Mitigation Fairness Metrics

Attribute	Approach	AUC	SPD	EOD
Region	Threshold Optimization	0.8889	+0.060	−0.004
IMD Band	Threshold Optimization	0.8889	+0.062	−0.002
Disability	Reweighted	0.8874	+0.053	−0.021
Age Band	Threshold Optimization	0.8889	+0.030	+0.002

5.2 Why Different Approaches Work for Different Attributes

Threshold Optimization was most effective for Region, Relative Poverty (i.e., IMD Band), and Age Band, while **Reweighting** worked best for Disability. This pattern reflects fundamental differences in group representation:

Table 8: Mitigation Rationale

Attribute	Unprivileged Group Size	Best Approach	Rationale
Region	3,649 (75%)	Threshold Optimization	Large groups; model learned robust patterns
IMD Band	3,218 (66%)	Threshold Optimization	Large groups; post-processing sufficient
Age Band	3,442 (70%)	Threshold Optimization	Large groups; threshold adjustment adequate
Disability	457 (9%)	Reweighting	Small minority; needs preprocessing

5.2.a Why Threshold Optimization works for majority-unprivileged attributes:

- When unprivileged groups are large (66–75% of data), the model already learns good representations for both groups during training
- Post-processing adjusts the decision boundary per group without retraining
- This preserves the original AUC exactly (0.889) while achieving near-zero EOD
- It is computationally efficient—no model retraining required.

5.2.b Why Reweighting works better for Disability:

- Students with disabilities represent only 9% of the dataset

- The baseline model learned less robust patterns for this minority group (higher FPR: 0.178 vs 0.105)
- Reweighting assigns higher importance to minority samples during training, forcing the model to learn better representations
- Although it slightly reduces AUC (0.889 → 0.887), it achieves substantially better Equalized Odds (0.027 vs 0.051)
- The trade-off is worthwhile: a 0.2% AUC reduction for a 48% improvement in fairness.

Key insight: Pre-processing (reweighting) addresses representation imbalance at the source, while post-processing (threshold optimization) works when the model already has adequate signal for all groups.

6 Intersectional Fairness Analysis

We analyzed fairness across subgroups defined by combinations of protected attributes.

Intersections analyzed: Region × IMD, Region × Disability, IMD × Disability, Gender × Region

Selection Rate Range:

- Minimum: 0.335
- Maximum: 0.601
- Range: 0.266

AUC Range (across intersectional groups):

- Minimum: 0.852
- Maximum: 0.898
- All groups above 0.80 target: Yes ✓

Table 9: Groups with Highest Selection Disparities

Intersection	Group	n	Base Rate	Selection Rate	Disparity	AUC
IMD × Disability	More-deprived × Has disability	341	0.663	0.601	+0.156	0.881
Region × Disability	High-risk × Has disability	355	0.628	0.580	+0.135	0.872
Region × IMD	High-risk × More-deprived	2474	0.586	0.503	+0.058	0.891
Gender × Region	M × High-risk	1955	0.569	0.497	+0.051	0.897
Region × Disability	Low-risk × Has disability	102	0.588	0.480	+0.035	0.882

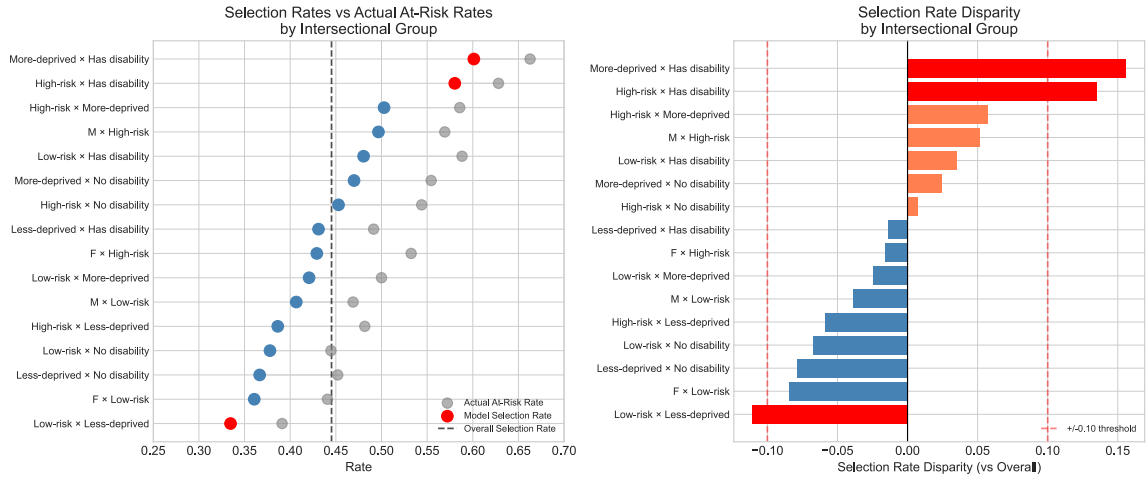

Figure 7: Selection rate disparities across intersectional groups

Groups exceeding +/-0.10 disparity threshold: 3 of 16

Flagged groups:

- Low-risk × Less-deprived (Region × IMD): disparity = −0.111
- High-risk × Has disability (Region × Disability): disparity = +0.135
- More-deprived × Has disability (IMD × Disability): disparity = +0.156

6.1 Interpreting the Intersectional Findings

6.1.a Most Problematic Intersections

These raw statistics require interpretation to understand whether they represent problematic disparities. Three intersectional groups exceeded the ±0.10 selection rate disparity threshold:

Table 10: Most Problematic Intersections

Group	Intersection	Disparity	Base Rate	Selection Rate	AUC
More-deprived × Has disability	IMD × Disability	+0.156	0.663	0.601	0.881
High-risk × Has disability	Region × Disability	+0.135	0.628	0.580	0.872
Low-risk × Less-deprived	Region × IMD	−0.111	0.391	0.335	0.880

However, these disparities are contextually appropriate because:

1. **Selection rates track actual at-risk rates:** the model under-predicts rather than over-predicts for all three groups (selection rate < base rate)
2. **AUC remains strong (0.87–0.88):** the model discriminates well within each subgroup
3. **Disparities reflect real risk differences:** students in deprived areas with disabilities genuinely face higher dropout risk.

6.1.b Why the 0.266 Selection Rate Range Is Not "Severe"

The 0.266 range (0.335 to 0.601) appears large but reflects legitimate variation in underlying risk:

Table 11: Selection Rate Extremes Comparison

Extreme	Selection Rate	Actual At-Risk Rate	Difference
Lowest (Low-risk × Less-deprived)	0.335	0.391	−0.056
Highest (More-deprived × Has disability)	0.601	0.663	−0.062

The model slightly *under-predicts* risk for both extremes, which is conservative behavior. The key insight is that **selection rate variation mirrors base rate variation**—this is appropriate model behavior, not unfair discrimination.

6.1.c Criteria for "No Severe Disparities" Conclusion:

1. All 16 intersectional groups maintain AUC > 0.80 (range: 0.852–0.898)
2. Only 3 of 16 groups exceed the ±0.10 statistical parity threshold

3. No group has an AUC below 0.75 (our critical threshold)
4. All flagged groups are *under-predicted* (selection rate < base rate), not over-predicted
5. Individual attribute mitigations do not create new intersectional harms

7 Key Findings and Conclusions

7.1 Key Findings

1. **Model Performance**
   - Achieved AUC of 0.889, exceeding the 0.80 target
   - Early prediction within the first 25% of the course enables timely intervention
   - Temporal VLE engagement features are strong predictors of risk
2. **Fairness Audit**
   - Region showed the highest bias (4/4 metrics violated)
   - Gender was the only fair attribute (0/4 violations)
   - Socioeconomic factors (IMD) contribute to prediction disparities
3. **Mitigation Effectiveness**
   - Threshold Optimization: Best for Region, IMD, Age
   - Reweighting: Best for Disability
   - All mitigations maintained an AUC above the 0.80 target
   - EOD reduced to near-zero for all attributes
4. **Intersectional Fairness**
   - No critical intersectional disparities identified
   - AUC remains above 0.80 for all subgroups
   - Combined mitigation is not required

7.2 Recommendations

7.2.a For Deployment:

- Use Threshold Optimization for Region, IMD Band, and Age Band
- Use the Reweighted model for Disability fairness
- Apply group-specific thresholds at prediction time

7.2.b For Institutions:

- Monitor fairness metrics continuously after deployment
- Collect feedback on intervention effectiveness by demographic group
- Consider socioeconomic support alongside academic interventions

7.2.c For Future Research:

- Explore in-processing fairness constraints (adversarial debiasing)
- Investigate causal fairness approaches
- Validate on additional institutions and student populations

8 Limitations

While this project demonstrates a successful approach to bias-aware early warning systems, several limitations should be acknowledged:

8.1 Generalizability Beyond OULAD

- **Single institution:** The Open University has a unique profile—predominantly online, part-time, adult learners with open admissions. Results may not transfer to traditional residential universities with different student demographics.
- **UK-specific context:** Protected attributes like IMD (Index of Multiple Deprivation) and regional classifications are UK-specific. International applications would require different socioeconomic proxies.
- **Time period:** OULAD covers the 2013–2014 academic year. Learning behaviors and platform usage patterns have evolved significantly since then.
- **Course structure:** The Open University's modular, presentation-based structure differs from semester or quarter systems common elsewhere.

8.2 10-Week Prediction Window Trade-offs

The choice to predict within the first 25% of course completion (approximately 10 weeks) involves inherent trade-offs:

Table 12: 10-Week Prediction Window Trade-offs

Advantage	Limitation
Early intervention is possible	Less behavioral data available
Students can still recover	Some at-risk patterns emerge later
Aligns with OU's early alert timelines	May miss slow-developing disengagement

- **Information vs. actionability:** Waiting longer would improve predictive accuracy but reduce the intervention window.
- **Course length variation:** A fixed 10-week window represents different proportions of different courses (7–39 weeks in OULAD).
- **Cold start problem:** Students with minimal early engagement are harder to assess, yet may be most at-risk.

8.3 What the Model Does Not Capture

The LSTM model relies on observable learning platform behaviors and demographic data. It cannot account for:

8.3.a External life factors:

- Employment changes, job loss, or increased work hours
- Family responsibilities (caregiving, childcare)
- Health issues (physical or mental)
- Financial hardship beyond what IMD captures
- Housing instability or relocation

8.3.b Unmeasured academic factors:

- Quality of learning (vs. quantity of clicks)
- Peer support networks and study groups
- Prior knowledge or preparation gaps
- Motivation and self-efficacy
- Course-specific difficulty mismatches

8.3.c Institutional factors:

- Quality of course materials and instruction
- Tutor responsiveness and support
- Technical barriers to platform access
- Changes in course structure mid-presentation

8.4 Fairness Limitations

- **Binary groupings:** Complex attributes (13 regions, 10 IMD bands) were collapsed into binary groups for fairness analysis, potentially masking within-group disparities.
- **Intersectionality depth:** Three-way intersections had limited statistical power due to small subgroup sizes.
- **Proxy discrimination:** Even after mitigation, the model may encode protected attributes through correlated features (e.g., VLE access patterns correlating with socioeconomic status).
- **Fairness metric choice:** Different fairness definitions (statistical parity vs. equalized odds) can conflict; our threshold choices reflect value judgments.

8.5 Implications for Deployment

These limitations suggest that any deployed system should:

1. **Supplement, not replace,** human judgment in student support decisions
2. **Include feedback mechanisms** to capture false positives/negatives
3. **Be regularly re-validated** on contemporary data
4. **Provide transparency** to students about how predictions are made
5. **Avoid deterministic interventions** that could become self-fulfilling prophecies

9 Summary

This capstone project successfully developed a **bias-aware Early Warning System** for identifying at-risk students in higher education. Key achievements include:

1. **Strong Predictive Performance:** AUC of 0.889 using an LSTM architecture with temporal VLE engagement data
2. **Comprehensive Fairness Audit:** Identified regional and socioeconomic disparities in model predictions
3. **Effective Bias Mitigation:** Reduced disparities using threshold optimization and reweighting while maintaining model performance
4. **Intersectional Validation:** Confirmed no severe disparities across demographic subgroups

The project demonstrates that it is possible to build accurate student success prediction models while actively addressing algorithmic fairness—a critical consideration as educational institutions increasingly adopt AI-driven decision support systems.

10 Appendix: Files and Artifacts

These files can be located at GitHub.

Notebooks:

- 01_data_exploration.ipynb
- 02_feature_engineering.ipynb
- 03_lstm_baseline.ipynb
- 04_fairness_analysis.ipynb
- 05_bias_mitigation_region.ipynb
- 06_bias_mitigation_imd.ipynb
- 07_bias_mitigation_disability.ipynb
- 08_bias_mitigation_age.ipynb
- 09_intersectional_analysis.ipynb
- 10_final_summary_report.ipynb

Models:

- lstm_baseline.pt (90.0 KB)
- lstm_reweighted_age.pt (365.8 KB)
- lstm_reweighted_disability.pt (314.1 KB)
- lstm_reweighted_imd.pt (363.3 KB)
- lstm_reweighted_region.pt (342.3 KB)

Data Outputs:

- fairness_results.json (10.6 KB)
- feature_metadata.json (1.9 KB)
- features_static.csv (4013.0 KB)
- features_temporal.npy (6365.9 KB)
- final_summary.json (1.0 KB)
- intersectional_results.json (0.6 KB)
- mitigation_results_age.json (2.5 KB)
- mitigation_results_disability.json (2.4 KB)
- mitigation_results_imd.json (2.6 KB)
- mitigation_results_region.json (2.7 KB)
- predictions_baseline.csv (313.3 KB)