LSAR's Report (verbatim)

Paper Information

- **Title**: Predicting** **Two-Year Public Versus Four-Year Postsecondary Enrollment** **from
- **Target Venue**: JEDM (confidence: 1.00)
- **Page Count**: 37 pages
- **Review Date**: 2026-08-29

1. Paper Summary (150-250 words)

The paper asks whether, among HSLS:09 students who ever attended postsecondary education, enrollment in a two-year public versus a four-year institution can be predicted from ninth-grade measures and high school transcript records. Using an analytic sample of 12,942 postsecondary entrants, the authors train five model families—logistic regression, elastic net, random forest, XGBoost, and stacking—and evaluate them on a school-aware held-out partition reconstructed from school-level variables because true school identifiers are suppressed. The primary result is moderate discrimination: XGBoost AUC = 0.704 (clustered CI [0.675, 0.730]), statistically indistinguishable from logistic regression (AUC = 0.699). Academic GPA dominates SHAP importance, followed by SES and ninth-grade math achievement. The model is well calibrated overall (Brier = 0.208; ECE = 0.031). Subgroup AUCs vary materially by race/ethnicity and SES, with gaps exceeding 10 percentage points. The paper claims three contributions: sector-specific prediction on HSLS:09, explicit calibration assessment, and subgroup fairness analysis. The authors interpret the modest accuracy as useful for screening but not deterministic assignment, and argue for subgroup-specific monitoring before deployment.

2. Relevance Assessment

- **venue_fit**: Strong — The paper addresses an educationally meaningful prediction target, uses established EDM methods, and explicitly engages fairness, calibration, and interpretability.
- **track_fit**: The paper fits EDM topics including equity, fairness, transparency; learner knowledge and performance modeling; learning analytics; and reproducibility. It also touches interpretability via SHAP.
- **scope_concerns**: No major out-of-scope flags. The work is not a methodological advance and does not introduce new data mining techniques, but as an applied predictive modeling study with fairness/calibration emphasis it is within scope for JEDM/EDM.

3. Strengths

- **Well-motivated outcome distinction**: The paper convincingly argues that "any postsecondary enrollment" and "two-year public versus four-year enrollment conditional on entry" are different prediction targets with different advising implications.
- **Calibration treated as first-class**: The paper reports Brier score, ECE, calibration slope, and calibration intercept, not just discrimination. This directly addresses a known underreporting problem in educational prediction.
- **Honest comparative finding**: The result that XGBoost and logistic regression are statistically indistinguishable is treated as substantive, not as failure, and leads to a practical

recommendation favoring the simpler model.
- **School-aware evaluation with clustered uncertainty**: The reconstructed pseudo-school split and clustered bootstrap represent a serious attempt to respect the nested HSLS:09 design, and the authors openly state the reconstruction is provisional.
- **Subgroup fairness promoted beyond an appendix**: Subgroup AUCs are reported for race/ethnicity and SES, with explicit thresholds for flagging disparities and appropriate caveats for small cells.

4. Weaknesses

- **MAJOR — Potential train/test leakage through preprocessing**: The text does not clearly state that median imputation, IterativeImputer, and one-hot encoding were fitted only on the training partition. If preprocessing was applied before the split, the reported AUCs may be optimistic and the model comparison biased.
- **MAJOR — Pseudo-school reconstruction cannot guarantee independent test schools**: Matching school-level profiles may merge distinct schools or split students from the same school across folds. The 927 reconstructed clusters versus an expected 944 is reassuring but not verification; true school leakage may remain, weakening the cross-school generalization claim.
- **MAJOR — Fairness analysis is incomplete relative to the stated research question**: Sex was specified as a protected attribute but could not be analyzed. The paper also reports subgroup discrimination only, not subgroup calibration or common fairness metrics such as equal opportunity, demographic parity, or threshold-specific false positive/negative rates.
- **MAJOR — No hyperparameter tuning or repeated validation described**: If tree/ensemble models used default settings, flexible models may not have been given a fair opportunity, potentially explaining logistic regression parity. A single train/test split also leaves the metric estimates sensitive to one partition.
- **MINOR — Missing-data handling is not fully reproducible**: The IterativeImputer configuration, seed, and exact sequence of median imputation versus complete-case encoding are not specified in enough detail for exact replication.

5. Detailed Dimensional Assessment

Dimension	Score (1-10)	Justification
Relevance	8	Predicting two-year versus four-year enrollment conditional on entry addresses a meaningful policy and advising question that is distinct from the more common any-enrollment prediction target, and it fits JEDM's focus on actionable educational data mining.

Dimension	Score (1-10)	Justification
Novelty	6	The outcome distinction and emphasis on calibration are useful contributions, but the overall modeling approach is fairly standard and the findings are largely confirmatory rather than introducing a new method or framework.
Theoretical/Conceptual Grounding	6	The paper is well motivated by sector-level differences in cost, completion, and transfer pathways, but the fairness framing is incomplete and the theoretical treatment of sector selection could be more deeply connected to prior postsecondary stratification research.
Methodological Rigor	4	Although the clustered bootstrap and calibration metrics show methodological awareness, major concerns about preprocessing leakage, inability to guarantee independent test schools in the pseudo-school reconstruction, and lack of hyperparameter tuning or repeated validation substantially weaken the rigor.
Empirical Support / Results	5	The paper reports useful discrimination, calibration, and subgroup results and treats the logistic regression/XGBoost parity honestly, but the validity of these estimates is threatened by potential leakage and reliance on a single train/test partition.

Dimension	Score (1-10)	Justification
Significance & Impact	6	The practical recommendation to prefer simpler logistic regression is valuable for adoption, and sector-specific prediction could inform advising, though the methodological concerns limit the strength of the conclusions.
Ethics, Fairness & Equity	6	Subgroup AUCs for race/ethnicity and SES with explicit thresholds are a strength, but the inability to analyze sex and the absence of calibration or threshold-based fairness metrics leave the equity analysis incomplete.
Clarity of Communication	7	The paper is generally clear about its goals, measures, and comparative findings, though missing preprocessing details and incomplete fairness reporting create some ambiguity for replication and interpretation.

6. Comparison with Related Work

Claimed contributions and novelty assessment

1. *First sector-specific prediction model for two-year public versus four-year enrollment using HSLS:09 with ninth-grade and transcript predictors.*
   Most relevant prior work: Alexander et al. (1987) examined two-year versus four-year college attendance using social background and academic determinants; He et al. (2025) predicted any college enrollment among low-SES HSLS:09 students; Plasman and Myles (2024) modeled four-year enrollment among students with learning disabilities.
   Assessment: Incremental. The substantive contrast is not new, and the dataset is already heavily used for enrollment prediction. The novelty lies in combining the exact outcome with this predictor set and ML evaluation.

2. *Explicit calibration assessment for sector prediction.*
   Most relevant prior work: Carriero et al. (2024) and Yang et al. (2024) demonstrated class-imbalance corrections can harm calibration; Tiruneh et al. (2025) reported logistic regression and XGBoost often have similar discrimination.
   Assessment: Incremental but useful. Applying these calibration lessons to a new educational outcome is a contribution, though the methods are not novel.

3. *Subgroup fairness analysis across sex, race/ethnicity, and SES.*

Most relevant prior work: Opoku et al. (2025) investigated accuracy–fairness trade-offs; Raftopoulos et al. (2025) compared bias mitigation strategies; Broden et al. (2026) emphasized group-conditional uncertainty.
Assessment: Incomplete relative to current fairness practice. The paper reports subgroup AUC but does not examine subgroup calibration or multiple fairness criteria. The contribution is therefore narrower than the framing suggests.

Overall, the paper is positioned as filling a gap between any-enrollment prediction and substantive sector-sorting research. That gap is real, but the paper's own novelty claims should be tempered. The comparative ML design is acceptable for an applied EDM study, especially because the transparent negative result and calibration emphasis add practical value beyond the individual model results.

Cited References

- K. Alexander, Scott Holupka, Aaron M. Pallas (1987). *Social Background and Academic Determinants of Two-Year versus Four-Year College Attendance: Evidence from Two Cohorts a Decade Apart*. American Journal of Education. DOI: 10.1086/443881 [VERIFIED] — Relevance: 8.0/10
- C. Belfield, P. Crosta (2012). *Predicting Success in College: The Importance of Placement Tests and High School Transcripts*. . DOI: 10.7916/D8K07CFB [VERIFIED] — Relevance: 5.0/10
- Michael Chajewski, K. Mattern, Emily J. Shaw (2011). *Examining the Role of Advanced Placement® Exam Participation in 4-Year College Enrollment*. . DOI: 10.1111/J.1745-3992.2011.00219.X [VERIFIED] — Relevance: 5.0/10
- Marcell Nagy, Roland Molontay (2023). *Interpretable Dropout Prediction: Towards XAI-Based Personalized Intervention*. International Journal of Artificial Intelligence in Education. DOI: 10.1007/s40593-023-00331-8 [VERIFIED] — Relevance: 7.0/10
- R. Guevara-Reyes, Iván Ortiz-Garcés, Roberto O. Andrade, Fernanda Cox-Riquetti, W. Villegas-Ch. (2025). *Machine learning models for academic performance prediction: interpretability and application in educational decision-making*. Frontiers in Education. DOI: 10.3389/feduc.2025.1632315 [VERIFIED] — Relevance: 5.0/10
- Z. Alamgir, Habiba Akram, S. Karim, Aamir Wali (2023). *Enhancing Student Performance Prediction via Educational Data Mining on Academic data*. Informatics Educ.. DOI: 10.15388/infedu.2024.04 [VERIFIED] — Relevance: 5.0/10
- N. Schmitt, Abigail Q Billington, Jessica Keeney, Ruchi Sinha (2009). *Alternatives to GPA 1 Prediction of Four-Year College Student Performance using Cognitive and Noncognitive Predictors and the Impact on Demographic Status of Admitted Students*. . https://www.semanticscholar.org/paper/aa7c41fc0e0627240ce13755e6b993a2b6d6bcc3 [VERIFIED] — Relevance: 5.0/10
- Bhimasen Moharana, Vinay Kumar Singh, T. Sarkar, Abass Hassan, Vikas Kumar Pandey (2024). *Predicting Indian Master's and PhD Enrollment using Data Mining & Machine Learning Approaches*. 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS). DOI: 10.1109/ICTACS62700.2024.10841260 [VERIFIED] — Relevance: 5.0/10

7. Questions for Authors

1. Were all imputation and encoding steps fitted only on the training partition? If not, please re-run the pipeline with train-only preprocessing and report whether the AUC and calibration metrics change materially.
2. How well does the pseudo-school reconstruction approximate true school membership? Can you provide evidence, such as cluster-size agreement beyond the 927-versus-944 total, or sensitivity analyses using alternative matching rules, to reassure readers that test schools are not contaminated?
3. Why was sex unavailable for the subgroup analysis, and can it be recovered by merging the protected-attribute file with the base-year student file? If not, the stated fairness contribution should be narrowed.
4. What hyperparameter tuning strategy was used for random forest, XGBoost, elastic net, and stacking? Were models selected via cross-validation or on the test set?

8. Suggestions for Improvement

1. **Guarantee leakage-free preprocessing**: Re-run the pipeline so all imputation, encoding, and scaling are learned exclusively on training folds. Report results both ways if necessary.
2. **Strengthen school-aware validation**: Add sensitivity analyses varying the clustering rule, and if possible use restricted-use data with true school IDs. If not, clearly narrow the generalization claim from "unseen schools" to "reconstructed school-like clusters."
3. **Complete and deepen fairness evaluation**: Include sex if possible, report subgroup-level calibration, and add threshold-based fairness metrics such as equal opportunity and false positive/negative rates. Report subgroup sizes and confidence intervals for all fairness quantities.
4. **Report hyperparameter tuning**: Specify search spaces, selection procedure, and nested cross-validation. If no tuning was performed, add a tuned comparison or explicitly acknowledge this limitation.
5. **Tone down "first" claims**: Either provide a reproducible search strategy for the retrieval or reframe the contribution as "to our knowledge" with explicit acknowledgment that sector-sorting itself is not new.

9. Minor Issues

- The stacking ensemble point estimate (AUC = 0.705) is slightly higher than XGBoost (0.704), but the text often calls XGBoost the "best model." Clarify this distinction or use "best individual model" consistently.
- The abstract and body state sex was part of the fairness analysis, but the results do not include sex. This should be flagged earlier or the research question narrowed.
- Some references cited in the Discussion, such as Glandorf et al. (2024), do not appear to be introduced in the Related Work section and may be missing from the reference list.
- The paper uses automated generation language in the author note; this is disclosed, but the authors should verify all cited references because several recent 2025–2026 citations could not be independently confirmed from the excerpt.
- Ensure all SHAP and calibration figures are legible and include axis labels and bin counts for ECE.

10. Overall Recommendation

- **Score**: 5.9 / 10
- **Recommendation**: Borderline
- **Confidence**: 5.1 - 6.7
- **Calibration (JEDM, n=10 published anchors)**: P25 5.30, median 5.5. This score is at or above the 25th percentile of published JEDM papers.
- **Measurement noise**: repeat reviews of an unchanged paper have differed by up to 1.9 points overall (measured test-retest MAD), and individual dimension scores by a full point on identical input. Do not read this score as precise to one decimal place.
