LSAR's Report (verbatim)

Paper Information

- **Title**: High** **School Preparation, Not Background, Predicts Postsecondary** **Persistence:
- **Target Venue**: JEDM (confidence: 1.00)
- **Page Count**: 39 pages
- **Review Date**: 2026-08-29

1. Paper Summary (150-250 words)

The paper uses HSLS:09 data to predict postsecondary persistence—remaining enrolled or having attained a credential by February 2016—among 12,918 students who ever attended postsecondary education. The authors train five models on pre-outcome high school transcript, expectation, and socioeconomic predictors, with a school-aware train/test split based on reconstructed pseudo-school clusters and clustered bootstrap confidence intervals. Logistic regression is the best-performing model (test AUC = 0.762), with good calibration. SHAP analysis suggests that high school GPA is the dominant predictor, while SES and ninth-grade achievement decline in importance relative to prior access models. Subgroup analyses show larger AUC differences by race than by SES quintile, with instability in the smallest racial subgroups. The paper argues that, conditional on enrollment, academic preparation matters more than background for persistence, and that deployment of such models would require subgroup-specific validation. The study is positioned as filling a gap between the access-prediction and persistence-prediction literatures.

2. Relevance Assessment

- **venue_fit**: Strong — The paper addresses a core educational data mining problem with a nationally representative dataset, predictive modeling, interpretability, and fairness analysis. It fits JEDM's scope for rigorous full-length empirical work and EDM topics such as learner knowledge/performance modeling, learning analytics, and equity/fairness.
- **track_fit**: It directly addresses "learner knowledge and performance modeling," "equity, privacy, transparency, fairness," and "learning analytics." It also touches "human factors, transparency, explainability" through SHAP.
- **scope_concerns**: The practical and policy claims are U.S.-centric and tied to HSLS:09. This is acceptable, but the generalizability claims should remain conditional on the U.S. postsecondary and transcript context. No other out-of-scope flags.

3. Strengths

- **Policy-relevant analytic sample selection**. The study restricts to students who ever attended postsecondary education, which directly addresses the conditional persistence question. This is a meaningful improvement over access models that predict enrollment on full cohorts and persistence models that already use institutional enrollees.
- **Attention to clustered data structure**. Despite suppressed school identifiers, the authors reconstruct pseudo-school clusters and use a school-aware train/test split and cluster bootstrap intervals. This shows appropriate concern for dependence structure and cross-school generalization.

- **Calibration and interpretability reported, not just discrimination**. The paper reports Brier score, ECE, calibration slope/intercept, SHAP values, partial dependence, and performance at a threshold. This goes beyond AUC-only reporting and supports practical interpretation.
- **Fairness reporting with some appropriate caution**. The study reports subgroup AUCs by race and SES, flags large racial gaps, and explicitly avoids substantive interpretation of tiny subgroups. It also reports SES quintile stability.
- **Transparent treatment of uncertainty and limitations**. The authors acknowledge the provisional nature of pseudo-school cluster reconstruction, the omission of sex subgroup analysis, and the uninterpretability of small racial subgroup estimates.

4. Weaknesses

- **MAJOR — The central access–persistence contrast is not empirically estimated in this paper.** The paper claims to show that SES and ninth-grade achievement dominate access models but recede in persistence models. However, no access model is trained on the same HSLS:09 data or even reported here. The contrast depends on "prior runs of this pipeline" and prior literature, not on a reproducible within-study analysis. This significantly weakens the central contribution.
- **MAJOR — Internal contradiction in feature importance claims.** The abstract states the next predictor after high school GPA is SES with mean $|\text{SHAP}| = 0.225$. But the Results section states race is the second-strongest grouped predictor with total SHAP = 0.338, followed by SES at 0.225. This is not a trivial wording issue; it affects the headline claim that "background" recedes. Race is a background/demographic variable and, if second, undermines the narrative.
- **MAJOR — Preprocessing leakage is not ruled out.** The Methods describe missing-data imputation and encoding before the train/test split, but do not state that imputers and encoders were fitted only on training data. Under a predictive modeling goal, this is a core validity concern. With 20.8% missingness in parental education and other imputed variables, full-data imputation could inflate test performance.
- **MAJOR — Model configuration and hyperparameter tuning are not described.** The paper concludes that logistic regression is the best model and that complex models do not improve over it, but if tree/ensemble models were not reasonably tuned, this comparison is not strong evidence. Reproducibility is limited by the absence of hyperparameter search details, software versions, seeds for all stochastic steps, and code/data availability.
- **MAJOR — Subgroup AUC point estimates are compared without confidence intervals.** Fairness conclusions rely on gaps exceeding a pre-registered 5-percentage-point threshold, but the larger within-cell gaps are not accompanied by bootstrap or analytic CIs. For fairness-sensitive inference, point estimates alone are insufficient, especially when sample sizes vary widely.
- **MINOR — Grouped SHAP may overstate categorical predictors.** Summing absolute SHAP contributions across one-hot encoded dummy columns gives variables with more levels, such as race, an inflated aggregate importance relative to single-column continuous variables. This may explain the race vs. SES discrepancy and should be corrected or explicitly discussed.

5. Detailed Dimensional Assessment

Dimension	Score (1-10)	Justification
Relevance	8	The paper addresses postsecondary persistence and inequality using a large national dataset, which is highly relevant to educational data mining and policy audiences. Restricting to students who ever enrolled is a meaningful, policy-relevant analytic choice.
Novelty	5	The modeling approach is largely an application of standard classifiers with fairness reporting. The central access-versus-persistence contrast is not empirically estimated within the same study, which reduces the novelty of the headline contribution.
Theoretical/Conceptual Grounding	6	The study is situated in access and persistence literature and uses relevant predictors, but the core theoretical contrast depends on prior work or prior pipeline runs rather than a fully integrated conceptual test.
Methodological Rigor	4	School-aware splitting and clustered bootstrap intervals are strengths, but major validity concerns remain: preprocessing leakage is not ruled out, hyperparameter tuning and reproducibility details are underspecified, and no code/data availability is described.

Dimension	Score (1-10)	Justification
Empirical Support / Results	4	Headline claims are weakened by an internal contradiction about whether SES or race is the second-strongest predictor, and subgroup AUC gaps are compared without confidence intervals, undermining statistical confidence in fairness conclusions.
Significance & Impact	5	The policy question has potential importance, but the current evidential gaps and replication concerns limit the likely impact on practice or policy.
Ethics, Fairness & Equity	6	The paper includes fairness reporting by race and SES with some appropriate caution, but the lack of confidence intervals for subgroup gaps and omission of sex subgroup analysis weaken the equity analysis.
Clarity of Communication	5	The paper is generally structured for an interdisciplinary audience, but contradictory feature-importance statements and incomplete method reporting create confusion about the main claims.

6. Comparison with Related Work

Claimed contributions and novelty assessment

1. **Claimed contribution**: Conditional ML prediction of postsecondary persistence using pre-college HSLS:09 transcript data.
   Most relevant prior work: Carballo-Mendívil et al. (2026), Tang et al. (2025), Cardona et al. (2020), and Waugh et al. (1994).
   Assessment: Incremental. The use of HSLS:09 transcript variables is valuable, but the predictive relevance of high school GPA for retention/dropout is well established. The novelty is mainly in the dataset and conditional sample, not in the predictor finding itself.

2. **Claimed contribution**: Contrasting access-model and persistence-model predictor rankings on the same analytic cohort.
   Most relevant prior work: Allensworth et al. (2026) for access; Tang et al. (2025) and

Carballo-Mendívil et al. (2026) for persistence/dropout.
Assessment: This is the paper's most novel intended contribution, but the manuscript does not actually implement an access model on the same HSLS:09 data. Without that, the contrast is only a comparison to previous pipeline outputs and literature. As presented, the claim is not adequately supported.

3. **Claimed contribution**: Subgroup fairness evaluation by race and SES.
Most relevant prior work: Verger et al. (2023), Raftopoulos et al. (2025), Gazi et al. (2026), and López-López et al. (2026).
Assessment: Reporting subgroup AUCs is useful, but this is becoming standard. The analysis would benefit from fairness metrics like MADD or equalized odds, plus confidence intervals for subgroup estimates.

Position relative to key work

The paper is strongest when it engages Carballo-Mendívil et al. (2026), which found high school GPA, parental education, and household assets among dominant dropout predictors. The present study partially replicates that finding but does not sufficiently distinguish its contribution beyond the national dataset and conditional enrollment sample.

The paper's claim that no retrieved study contrasts access and persistence predictor rankings on the same analytic cohort is plausible, but the response should have been to conduct that contrast directly. As written, the contrast remains unverified.

Baselines are appropriate as model families, but the model comparison is weak because hyperparameter tuning is not described. The inclusion of ElasticNet, random forest, XGBoost, and stacking is reasonable, but the conclusion that logistic regression is best may reflect default settings rather than genuine model family superiority.

Cited References

- Gordon W. Waugh, T. Micceri, P. Takalkar (1994). *Using Ethnicity, SAT/ACT Scores, and High School GPA To Predict Retention and Graduation Rates..* . https://www.semanticscholar.org/paper/c26 [VERIFIED] — Relevance: 9.0/10
- Juntao Chen, Xiaodeng Zhou, Jiahua Yao, Su-Kit Tang (2025). *Application of machine learning in higher education to predict students' performance, learning engagement and self-efficacy: a systematic literature review.* Asian Education and Development Studies. DOI: 10.1108/aeds-08-2024-0166 [VERIFIED] — Relevance: 8.0/10
- T. Cardona, E. Cudney, R. Hoerl, Jennifer Snyder (2020). *Data Mining and Machine Learning Retention Models in Higher Education.* Journal of College Student Retention. DOI: 10.1177/1521025120964920 [VERIFIED] — Relevance: 8.0/10
- M. Verger, Sébastien Lallé, F. Bouchet, Vanda Luengo (2023). *Is Your Model "MADD"? A Novel Metric to Evaluate Algorithmic Fairness for Predictive Student Models.* Educational Data Mining. DOI: 10.5281/zenodo.8115786 [VERIFIED] — Relevance: 9.0/10
- George Raftopoulos, Gregory Davrazos, S. Kotsiantis (2025). *Evaluating Fairness Strategies in Educational Data Mining: A Comparative Study of Bias Mitigation Techniques.* Electronics. DOI: 10.3390/electronics14091856 [VERIFIED] — Relevance: 8.0/10

7. Questions for Authors

1. The analytic sample is reported as 12,918 students, but the stated persister and non-persister counts sum to 10,289. Are 8,219 persisters and 2,070 non-persisters training-set counts? Please report the full sample counts separately for training and test, along with the exact non-persistence rate in each split.
2. How were imputation, encoding, scaling, and any other preprocessing steps handled relative to the train/test split? Were all imputers and encoders fitted on training data only and then applied to the test set, or were they fitted on the full sample before splitting?
3. Why does the abstract identify SES as the second-strongest predictor when the Results section states race has higher grouped SHAP importance? Was race excluded from the "background" narrative, and if so, what justifies that exclusion?
4. Why was an access model not trained on the same HSLS:09 analytic cohort in this study? Without that within-study access model, the central access–persistence contrast appears to rely on prior pipeline outputs rather than an empirical comparison in the current paper.

8. Suggestions for Improvement

1. **Add a direct access model comparison.** Train an access/persistence outcome model on the full HSLS:09 ninth-grade cohort, or at least on a comparable pre-enrollment sample, and report the same SHAP/importance and subgroup analyses. This would make the central contrast reproducible and internal to the paper.
2. **Correct and clarify feature importance reporting.** Do not sum absolute SHAP values across one-hot encoded dummies without accounting for variable dimensionality. Report grouped importance using a method that makes categorical and continuous variables comparable, or clearly report both raw and degree-of-freedom-adjusted contributions.
3. **Document preprocessing as leakage-free and provide reproducibility details.** Specify that all imputation, scaling, and encoding were fit only on training data. Include hyperparameter search ranges, software package versions, random seeds, and ideally code/data access or a detailed run log.
4. **Add confidence intervals to subgroup AUCs.** For all subgroup performance estimates, especially racial subgroups with n > 100, report bootstrap or analytic CIs. Use inferential comparisons rather than only a 5-percentage-point gap threshold.
5. **Temper causal and policy language.** The model is predictive, not causal. Reword statements such as "support should target academic preparation and first-year transition rather than financial aid alone" to reflect that these are predictive associations and require causal or implementation research before policy translation.

9. Minor Issues

- The abstract reports the best-model CI as [0.740, 0.785], while the Results text reports [0.740, 0.784].
- The sentence in Section 2/fairness contribution paragraph contains garbled text: "*mldesigntopostsecondarypersistenceconditionalonenrollmentusinganationallyrepresentativelo*" and related corrupted phrases. This must be fixed.
- The claim that roughly one in three enrollees leaves without a credential lacks a citation in the introduction.
- Table/Figure references are present, but the figure/table content is not available in the review excerpt; verify that all tables and figures match the text.

- The phrase "more than twice the SHAP importance of the next predictor" conflicts with the reported race total SHAP of 0.338. This should be corrected wherever it appears.
- X1PAREDU missingness is reported as exceeding 20%, but the imputation/robustness discussion should state whether the sensitivity analysis changed any subgroup or fairness conclusions, not only AUC and top-5 features.

10. Overall Recommendation

- **Score**: 5.2 / 10
- **Recommendation**: Borderline
- **Confidence**: 4.4 - 6.1
- **Calibration (JEDM, n=10 published anchors)**: P25 5.30, median 5.5. This score is below 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.

Disclaimer: This review was generated by LSAR (Learning Science Auto-Reviewer), an AI-assisted review system. It is intended to provide rapid, constructive feedback to help authors improve their work. It should NOT be used as a substitute for human peer review, nor should it be used in any way that violates the reviewing policies of the target venue. AI-generated reviews may contain errors, miss nuances, and have systematic biases (e.g., underweighting novelty, overweighting technical validity). Authors should use this feedback as one input among many. Conference reviewers should NOT use this tool to generate or supplement their official reviews, as this violates the policies of AIED, EDM, L@S, and LAK.