Engagement and Belonging Add Predictive Signal for High School Dropout Episodes Beyond Achievement and SES: Evidence from a Nationally Representative U.S. Cohort

EDM-ARS

Automated Research System

Author Note

Note: This manuscript was generated by EDM-ARS, an automated educational data mining research system, under human oversight. **Disclosures statement:** The authors declare no competing interests. **Funding:** No external funding was received.

Abstract

Early-warning systems aim to identify ninth graders at risk of a later dropout episode, yet most predictive work relies on administrative or single-institution data and treats non-completion rather than the dropout episode itself. Using the High School Longitudinal Study of 2009 (HSLS:09; N = 23,503), we tested whether fall-of-ninth-grade engagement and school-belonging measures add predictive value for a dropout episode by the 2013 update above and beyond achievement and socioeconomic status. A school-aware train/test split with clustered bootstrap confidence intervals yielded a best-model AUC of 0.73 (95% CI [0.71, 0.75]) for XGBoost. The engagement/belonging/expectations/school-context block added 0.060 AUC points (95% CI [0.045, 0.075]) over an achievement-plus-SES baseline, indicating independent signal. At an operating threshold flagging roughly two-thirds of true dropout episodes, the model produced 0.58 precision and 0.67 recall; the 11% base rate imposes a hard ceiling on precision that accuracy obscures. Subgroup analyses showed stable performance by sex but meaningful variation by race/ethnicity, with small-cell estimates flagged as unreliable. These results suggest that brief ninth-grade engagement and belonging surveys can improve risk stratification for dropout episodes at modest cost, provided predicted probabilities are recalibrated before threshold-based use.

Keywords: dropout prediction, early warning systems, school engagement, school belonging, HSLS:09, machine learning

Engagement and Belonging Add Predictive Signal for High School Dropout Episodes Beyond Achievement and SES: Evidence from a Nationally Representative U.S. Cohort

Introduction

Every fall, roughly four million students enter ninth grade in the United States, and a small but consequential fraction of them will experience a dropout episode before completing high school. The consequences of these episodes are well documented: interrupted schooling is associated with lower lifetime earnings, poorer health, and reduced civic participation, and the costs concentrate among students from lower-income families and historically marginalized racial and ethnic groups. Early-warning systems attempt to identify at-risk students early enough for intervention, and a large machine-learning literature now promises to improve such systems by mining administrative, survey, and learning-management data for predictive signal (Rodriguez-Ortiz et al., 2025; Shi et al., 2025). Yet the practical question facing a ninth-grade counselor is narrower than the literature's headline claims suggest: given only the information available in the fall of ninth grade, how well can we identify which students will later experience a dropout episode, and which pieces of that fall information actually carry the signal?

Two gaps separate the existing evidence base from that practical question. First, most dropout-prediction studies use administrative or learning-management data from a single institution or a single national context, and they typically model non-completion rather than the dropout episode itself (Barboza et al., 2026; Carballo-Mendívil et al., 2025; Fernandez et al., 2025; Pai et al., 2026). A dropout episode is a distinct outcome: a student may leave school and later return, and the intervention window for a student at risk of an episode differs from the window for a student at risk of permanent non-completion. Studies that conflate the two outcomes cannot tell a school whether its early-warning model is tuned for the right target. Second, although school engagement and school belonging occupy a central place in dropout theory, almost no predictive work isolates their incremental contribution above and beyond achievement and socioeconomic status. Engagement and belonging are correlated with achievement and SES; whether they carry independent signal for dropout episodes once those controls are present is an open empirical question, and it is the question that determines whether schools should invest in collecting brief engagement surveys rather than relying on test scores and demographic records alone.

The theoretical case for engagement and belonging as dropout predictors is strong. Engagement, defined as the behavioral, emotional, and cognitive investment a student makes in school, declines across adolescence on average (Symonds et al., 2026), and low engagement is a consistent correlate of withdrawal from schooling. Belonging, the felt sense of being accepted and valued in the school community, is similarly implicated in persistence. Both constructs are measured in the High School Longitudinal Study of 2009 (HSLS:09), a nationally representative U.S. cohort that followed ninth graders through high school and beyond, and both are available in the fall of ninth grade, before most dropout episodes occur. HSLS:09 therefore offers a rare opportunity to test whether these constructs add predictive value in exactly the setting where an early-warning system would operate: a nationally representative sample, with fall-of-ninth-grade information only, predicting a later dropout episode.

This study uses HSLS:09 to answer three linked questions. First, how accurately can a dropout episode by the 2013 update be predicted from fall-2009 ninth-grade information alone? Second, how much of that accuracy comes from engagement and school-context measures rather than prior achievement and socioeconomic status? Third, how does the precision-recall trade-off behave at the operating thresholds an early-warning system would actually use? The third question matters because the outcome is rare: approximately 11% of HSLS:09 ninth graders experienced at least one dropout episode by the 2013 update. At that base rate, a model can achieve high accuracy by predicting that no student drops out, which is useless for intervention. The operationally relevant metrics are precision and recall at explicit flagging thresholds, and these are precisely the metrics that most machine-learning-for-dropout papers do not report (Liu et al., 2025; Setiawan et al., 2025).

We make three contributions. First, we provide what is, to our knowledge, the first nationally representative estimate of the incremental predictive value of ninth-grade engagement and school-belonging measures for a dropout episode, above and beyond achievement and socioeconomic status. We estimate this increment directly by comparing a model restricted to achievement and SES against a full model that adds engagement, belonging, expectations, and school context, and we report the difference with a cluster-bootstrap confidence interval. Second, we report precision-recall behavior at explicit early-warning operating thresholds rather than treating accuracy or AUC as the headline metric. We show what a school would actually observe if it flagged students above a given risk score: how many true dropout episodes it would catch, and how many false alarms it would generate. Third, we evaluate all models under a school-aware train/test split, so that the headline test metric is computed over students in schools never seen during training, and we report clustered confidence intervals that account for within-school correlation. We complement this with subgroup analyses by sex, race/ethnicity, and socioeconomic status, and with a sensitivity analysis that drops the four predictors with the highest missingness.

The findings are substantively clear. The engagement and belonging block adds roughly six AUC points beyond achievement and SES alone, and the confidence interval excludes zero; engagement and belonging are not merely proxies for prior achievement. At the same time, the best model's discrimination is modest (AUC approximately 0.73), and its calibration is poor, which means that predicted probabilities should not be taken at face value at decision thresholds. The precision-recall analysis shows that flagging roughly two-thirds of true dropout episodes generates about 1.7 false positives per true positive, a trade-off that schools must weigh explicitly. The subgroup analyses reveal essentially no sex gap but substantial variation across racial/ethnic groups, with the model less informative for Black and Asian students than for White students, a pattern that carries fairness implications for early-warning deployment.

The remainder of the paper proceeds as follows. Section 2 reviews the dropout-prediction literature, the theoretical grounding for engagement and belonging, and the early-warning-systems literature on operating thresholds. Section 3 describes the HSLS:09 data, the predictor blocks, the school-aware split, and the model battery. Section 4 reports the block-incremental analysis, the model comparison, the precision-recall trade-off, calibration, school-level variance, subgroup performance, and the sensitivity analysis. Section 5 interprets these findings for early-warning system design, and Section 6 states the limitations and concrete directions for future work.

Related Work

Dropout Prediction from Administrative and Survey Data

Machine learning approaches to dropout prediction have matured rapidly over the past decade, but the field remains dominated by studies that draw on administrative records, learning management system (LMS) traces, or single-institution datasets. A systematic review of 161 studies published between 2014 and 2024 found that most at-risk prediction work relies on behavioral engagement and academic factors extracted from online learning platforms, with deep learning and ensemble methods increasingly displacing classical models (Shi et al., 2025). The review also documented persistent concerns about reproducibility, generalizability, and interpretability across the literature. A second systematic review focused on learning analytics in higher education reached a similar conclusion: engagement prediction and dropout-risk modeling dominate the field, but generative AI applications remain experimental and face unresolved transparency and pedagogical-grounding problems (Rodriguez-Ortiz et al., 2025).

The geographic and institutional concentration of this work is striking. Studies of dropout prediction in East African secondary schools (Pai et al., 2026), Peruvian public schools (Barboza et al., 2026), Mexican public universities (Carballo-Mendívil et al., 2025), Colombian educational systems (Fernandez et al., 2025), and Indian high schools (Nagar et al., 2026) each report high accuracy or AUC on their respective datasets, but each is bound to a single national or institutional context. The East African study, for instance, achieved an AUROC of 0.982 on a sample of 62,739 students, with SHAP analysis identifying household size and transportation barriers as primary risk factors (Pai et al., 2026). The Peruvian study reported 97.27% accuracy with gradient boosting on 1,456 records from a single metropolitan school (Barboza et al., 2026). These results are impressive within their contexts, but they do not address the question of whether the same predictive structure holds in a nationally representative cohort with a different outcome definition.

A second limitation of the existing literature is the outcome variable itself. Most studies model non-completion, attrition, or course-level failure rather than the dropout episode as a distinct event. The systematic review by Shi et al. (2025) identified four primary at-risk definitions in the reviewed studies, each tied to specific stages of engagement and performance in a course, but none corresponding to a discrete dropout episode recorded in a longitudinal survey. The higher-education literature is similarly focused on program-level attrition: Liu et al. (2025) predicted dropout, enrollment, and graduation as three separate classes in a Portuguese university dataset, while Sulak and Koklu (2025) used the same three-class scheme with decision trees, random forests, and neural networks. The distinction matters because a dropout episode and non-completion are not the same event: a student may experience a dropout episode and later re-enroll and complete, a pattern that is invisible to models trained only on final completion status. The present study uses X3EVERDROP, which records at least one dropout episode by the 2013 update, and therefore speaks to a different, and arguably more actionable, target for early-warning systems.

The predictor sets used in prior work also differ systematically from the one examined here. Administrative and LMS-based studies emphasize attendance, course performance, and behavioral traces (Alzahrani et al., 2025; Hung et al., 2025; Kubsch et al., 2025). Survey-based studies, where they exist, tend to use demographic and socioeconomic variables rather than psychological constructs like school belonging or engagement (Carballo-Mendívil et al., 2025). The K-12 online learning case study by Hung et al. (2025) identified five at-risk types through interpretable AI, but the predictors were behavioral traces from a learning platform, not self-reported engagement or belonging. The self-regulated learning study by Kubsch et al. (2025) used cognitive, metacognitive, and affective data from digital workbooks, again without the school-belonging construct that is central to the present study. The gap is therefore twofold: no retrieved study uses a nationally representative U.S. sample, and no retrieved study isolates the incremental contribution of engagement and belonging measures over achievement and socioeconomic status.

Engagement, Belonging, and Dropout Theory

The theoretical case for engagement and belonging as dropout predictors is well established in the educational psychology literature, even though the machine learning literature has been slow to incorporate these constructs. Stage-environment fit theory proposes that student engagement declines across adolescence as school environments increasingly fail to meet adolescents' developing psychological needs (Symonds et al., 2026). A meta-analysis of 125 longitudinal studies confirmed a general decline in engagement across adolescence, with the steepest declines occurring in early adolescence and during school transitions (Symonds et al., 2026). This developmental trajectory is directly relevant to the present study, which measures engagement and belonging in the fall of ninth grade, precisely the period when engagement is declining most sharply and when dropout risk begins to crystallize.

School belonging, the second focal construct, has a parallel theoretical lineage. Belonging captures the student's sense of being accepted, respected, and supported within the school community, and it is conceptually distinct from engagement, which captures behavioral and emotional investment in learning. The two constructs are correlated but not interchangeable, and both are hypothesized to operate through mechanisms that are partially independent of achievement: a student may perform adequately on standardized tests yet feel disconnected from school in ways that elevate dropout risk. The psychosocial development literature reinforces this point, showing that engagement and related constructs decline from early to late adolescence and differ systematically by gender (de Carvalho et al., 2025). The longitudinal study by Heikkinen et al. (2025) found that engagement and self-regulation develop together over time, with distinct behavioral patterns that have direct implications for intervention design.

The critical empirical question, however, is whether these constructs carry independent predictive signal once achievement and socioeconomic status are controlled. Engagement and belonging are correlated with both: students from higher-SES families and students with higher test scores tend to report higher engagement and belonging. If the correlation is strong enough, engagement and belonging measures may add little to a model that already includes achievement and SES, making their collection in early-warning systems redundant. The universal behavior screening study by Graybill et al. (2025) is one of the few to address this question directly, finding that self-reported emotional and behavioral symptoms predicted early-warning indicators after controlling for fall disciplinary incidents, and that these symptoms accounted for some of the variance previously attributed to race and gender. That study, however, used middle school students and a different outcome set, leaving the question open for ninth-grade dropout episodes in a nationally representative cohort. The present study addresses this gap by fitting a block-incremental model that compares achievement-plus-SES alone against the full predictor set, isolating the marginal contribution of the engagement, belonging, expectations, and school-context block.

Early-Warning Systems and Operating Thresholds

Early-warning systems in practice do not report AUC values to school counselors; they flag students for intervention based on a risk threshold. This operational reality makes precision and recall at specific thresholds the metrics that matter, yet the machine learning literature on dropout prediction has been slow to adopt this framing. The systematic review by Shi et al. (2025) found that most studies report accuracy, AUC, or F1 as headline metrics, with little attention to the threshold at which a model would actually be deployed. The higher-education study by Carballo-Mendívil et al. (2025) is a partial exception, reporting sensitivity of 88% for the dropout class, but it does not present the full precision-recall trade-off at multiple operating points. The graduate-program study by Syaripudin et al. (2026) calibrated a cost-sensitive alert threshold and reported recall of 0.613 at that threshold, but the study used synthetic data and did not address the base-rate problem that dominates real-world dropout prediction.

The base-rate problem is the central challenge for early-warning systems targeting rare events. When only 11% of students experience a dropout episode, a model with AUC 0.73 cannot achieve high precision and high recall simultaneously: flagging two-thirds of true dropouts necessarily generates a substantial number of false positives. The class-imbalance literature has addressed this problem primarily through resampling techniques like SMOTE (Putra & Utami, 2025; Setiawan et al., 2025), but resampling changes the training distribution without changing the fundamental precision ceiling imposed by the base rate. The chronic absenteeism study by Wu and Weiland (2026) demonstrated that SMOTE combined with XGBoost can substantially improve recall over logistic regression, but it did not report the precision cost of that improvement. The present study takes a different approach: rather than resampling, it reports precision, recall, F1, and F2 at explicit operating thresholds, making the trade-off visible and letting practitioners decide what false-positive rate is acceptable for their intervention budget.

The fairness dimension of threshold-based deployment has also received increasing attention. Opoku et al. (2025) examined accuracy-fairness trade-offs in student performance prediction using the Open University Learning Analytics Dataset, finding that standard models exhibit bias across demographic subgroups and that mitigation techniques can reduce disparities while maintaining acceptable accuracy. The East African study by

Pai et al. (2026) reported minimal bias across gender and language groups, but its fairness analysis was limited to aggregate metrics rather than subgroup-specific performance. Neither study addressed the interaction between operating thresholds and subgroup fairness: a threshold that is optimal for the overall population may produce systematically different flagging rates across racial or ethnic groups. The present study reports subgroup-specific AUCs for sex, race/ethnicity, and SES, and interprets the observed gaps in light of the small sample sizes that characterize minority subgroups in the test set.

The interpretability literature has similarly emphasized the importance of explaining model predictions to practitioners, but it has not connected explanation quality to threshold-based decision-making. Liu et al. (2025) used SHAP to identify curricular units completion as the dominant predictor in their ensemble model, while Syaripudin et al. (2026) argued that explanation quality, rather than raw discrimination, is the decisive ingredient for adoption of dropout early-warning systems. The present study follows this tradition by reporting SHAP values for the best-performing model, but it goes further by connecting the SHAP results to the block-incremental analysis, showing which constructs contribute independently to prediction and which are redundant with achievement and SES.

In sum, the retrieved literature establishes that machine learning can predict dropout with moderate to high accuracy in single-institution and developing-country contexts, that engagement and belonging are theoretically grounded dropout predictors whose independent signal remains untested in a nationally representative U.S. cohort, and that early-warning systems operate at thresholds where precision-recall trade-offs, not aggregate accuracy, determine practical utility. The present study addresses all three gaps simultaneously: it uses a nationally representative U.S. sample, it isolates the incremental contribution of the engagement and belonging block over achievement and SES, and it reports precision-recall behavior at explicit operating thresholds, with subgroup fairness checks and clustered confidence intervals that account for the multilevel structure of the data.

Methods

Data and Sample

We used the public-use file of the High School Longitudinal Study of 2009 (HSLS:09), a nationally representative cohort study conducted by the National Center for Education Statistics. HSLS:09 followed a stratified two-stage probability sample of ninth-grade students from 944 schools, with approximately 25 students sampled per school. The analytic sample comprised all 23,503 ninth graders with a known dropout status at the 2013 update wave. The outcome, X3EVERDROP, records whether a student experienced at least one dropout episode by the 2013 follow-up; 2,576 students (11.0%) had at least one such episode. We emphasize that X3EVERDROP captures a dropout episode, not non-completion: a student who left school and later re-enrolled or earned a credential is counted as a positive case. This distinction matters for comparability with studies that model non-completion as the outcome (Barboza et al., 2026; Pai et al., 2026).

Because school identifiers are suppressed in the public-use file, we reconstructed school clusters by grouping students with matching school-level variable profiles (school climate scale, school control, locale, and region). This yielded 935 pseudo-school clusters against an expected 944 based on the HSLS:09 sampling frame, with a mean cluster size of 25.1 students (median 25, range 1–96). The reconstruction cannot be verified against true school membership, so any claim of generalization to unseen schools is provisional on the reconstruction approximating true school membership. We used these reconstructed clusters to create a school-aware train/test split via StratifiedGroupKFold, assigning 18,806 students from 751 clusters to training and 4,697 students from 184 clusters to testing, with zero school overlap between partitions. This split ensures that the headline test metric is computed over students in schools never seen during training, providing a cross-context generalization estimate rather than a within-school interpolation estimate.

Predictors and Missing Data

We selected 14 predictors from the fall-2009 ninth-grade wave, organized into five conceptual blocks. The achievement and socioeconomic status block comprised X1TXMTSCOR (ninth-grade standardized mathematics achievement) and X1SES (a continuous socioeconomic status composite). The engagement and belonging block comprised X1SCHOOLENG (school engagement scale) and X1SCHOOLBEL (sense of school belonging scale). The expectations block comprised X1STUEDEXPCT (student educational expectations) and X1PAREDEXPCT (parent educational expectations for the student). The family background block comprised X1PAREDU (parents' highest education). The school context block comprised X1SCHOOLCLI (administrator-reported school climate), X1LOCALE (school urbanicity), X1CONTROL (school sector), and X1REGION (census region). Demographic predictors were X1DUALLANG (dual first-language indicator), X1SEX, and X1RACE. All predictors were measured at the base-year wave, temporally prior to the 2013 outcome, so no leakage from later waves was possible.

Missingness was non-trivial for several predictors. Three variables exceeded 28% missingness: X1PAREDEXPCT (35.7%), X1PAREDU (28.4%), and X1STUEDEXPCT (28.6%). One additional variable, X1SCHOOLCLI, had 20.6% missingness. The remaining predictors had missingness below 13%, and X1LOCALE, X1CONTROL, and X1REGION were complete. We imputed all missing values with IterativeImputer, a multivariate imputation strategy that models each variable with missing values as a function of the others in a round-robin fashion. For categorical variables with minimal missingness (X1LOCALE, X1CONTROL, X1REGION, X1SEX, X1RACE), we used mode imputation. The outcome was never imputed. Because three predictors exceeded the 20% missingness threshold at which imputation choices can meaningfully influence findings, we conducted a sensitivity analysis (reported in the Results) that refit the best-performing model class after excluding all four high-missingness variables.

Models and Evaluation

We fit a full model battery of five classifiers: logistic regression, random forest, XGBoost, an elastic-net regularized linear model, and a stacking ensemble that combined the four individual models with a logistic meta-learner. All models were tuned with StratifiedGroupKFold cross-validation (5 splits, shuffle, random seed 42), which preserves both class proportions and school-group separation within each fold. Hyperparameter grids were modest: for random forest, we searched over n_estimators in {100, 300, 500}, max_depth in {5, 10}, and min_samples_leaf in {1, 5, 10}; for XGBoost, over learning_rate in {0.01, 0.05, 0.1}, n_estimators in {100, 300, 500}, and max_depth in {3, 5, 7}; for elastic net, over alpha in {0.001, 0.01, 0.1, 1.0} and l1_ratio in {0.1, 0.5, 0.7, 0.9}. All tuning used ROC-AUC as the scoring metric. The best hyperparameters were max_depth 5, min_samples_leaf 10, n_estimators 500 for random forest; learning_rate 0.05, max_depth 3, n_estimators 300 for XGBoost; and alpha 0.001, l1_ratio 0.9 for elastic net.

The primary evaluation metric was area under the receiver operating characteristic curve (AUC), chosen because the 11% base rate makes accuracy a misleading summary of model quality. We computed 95% confidence intervals for every model's AUC using two bootstrap procedures. The standard row-level bootstrap resampled test-set rows with replacement (1,000 iterations, seed 42). The clustered bootstrap resampled whole school clusters with replacement, then collected every row in the sampled clusters, preserving within-school correlation. Because the intraclass correlation for the outcome was 0.0615 (small but non-negligible by conventional thresholds), the clustered intervals are the appropriate uncertainty statements for the primary metric; we report both so the difference is auditable.

We conducted a block-incremental analysis to answer the study's central question: how much predictive signal do the engagement, belonging, expectations, and school-context measures add beyond achievement and SES alone? We fit a reduced model using only

X1SES and X1TXMTSCOR, then compared its test-set AUC to the full model's AUC using a cluster-aware bootstrap difference test (1,000 replicates). The difference and its confidence interval quantify the incremental contribution of the non-achievement block.

We also ran a moderation analysis to test whether the incremental contribution of the engagement/belonging block varied by socioeconomic status. We fit the full model and a model with interaction terms between the engagement/belonging predictors and X1SES, then compared them with a likelihood-ratio test. We additionally computed the engagement/belonging block's incremental AUC within tertiles of X1SES and bootstrapped the difference between the top and bottom tertiles.

To characterize the precision-recall trade-off at operating thresholds an early-warning system would actually use, we fit two variants of the best-performing model class: one with balanced class weights and one with no class weighting. We report precision, recall, F1, and F2 for both variants, along with the confusion matrix at the default 0.5 threshold. F2 weights recall twice as heavily as precision, reflecting the early-warning use case in which missing a true dropout episode is more costly than flagging a false positive.

We quantified calibration with the Brier score, expected calibration error (ECE) across 10 probability bins, and the calibration slope and intercept from a logistic recalibration of predicted probabilities against observed outcomes. A discriminative model can still be badly calibrated, and calibration matters directly for threshold-based decision-making.

For interpretability, we computed SHAP values for the best individual model (XGBoost) using TreeExplainer. We report mean absolute SHAP values as feature importance and present SHAP summary and importance plots, along with partial dependence plots for the three most important predictors. We note that SHAP directions for individual predictors are interpreted relative to the model's reference behavior, and collinearity among correlated predictors (e.g., achievement, SES, and parent expectations) can produce counterintuitive directions that should not be read as causal effects.

We did not apply synthetic minority oversampling (SMOTE). The school-aware train/test split cannot assign synthetic rows to school groups, and oversampling within training folds would break the group structure that the clustered evaluation depends on. Class imbalance was instead handled through threshold-based and rank-based metrics (AUC, precision-recall behavior, F2) and through the balanced class-weight variant of the best model.

Subgroup performance was evaluated for the three protected attributes specified in the research design: X1SEX, X1RACE, and X1SES. For each attribute, we computed the test-set AUC separately for each level (for X1SES, a continuous variable, we used quintile bins). We flagged any attribute whose maximum minus minimum subgroup AUC exceeded 5 percentage points as a fairness concern, and we documented small-n subgroups (fewer than 100 students) as potentially unreliable.

This study was conducted using EDM-ARS, an automated educational data mining research system. All data preparation, model training, evaluation, and manuscript generation were performed programmatically.

Results

Engagement and Belonging Add Meaningful Incremental Signal Beyond Achievement and SES

The central question is whether ninth-grade engagement and school-belonging measures carry predictive signal for a later dropout episode that is not already captured by achievement and socioeconomic status. We answered this with a block-incremental analysis: a baseline model using only X1TXMTSCOR (ninth-grade math achievement) and X1SES (socioeconomic status composite) was compared against the full 14-predictor model. The baseline achieved an AUC of 0.669, while the full XGBoost model achieved 0.730, a difference of 0.060 (95% CI [0.045, 0.075], cluster bootstrap, 1,000 replicates). The confidence interval excludes zero, so the engagement, belonging, expectations, and school-context block adds roughly six AUC points beyond achievement and SES alone.

This increment is not merely a re-expression of achievement. The SHAP analysis for the full model ranks X1SCHOOLENG (school engagement, mean |SHAP| = 0.175) and X1SCHOOLBEL (school belonging, mean |SHAP| = 0.138) fourth and fifth in importance, behind X1TXMTSCOR (0.374), X1SES (0.250), and X1PAREDEXPCT (0.179). Figure 1 shows the grouped feature importance, and Figure 2 shows the distribution of SHAP values across the test set. The engagement and belonging scales are correlated with achievement and SES, but their SHAP contributions persist after those controls are in the model, which is the operational definition of independent signal in a tree ensemble.

The direction of the engagement and belonging effects is consistent with dropout theory: lower engagement and lower belonging are associated with higher predicted dropout risk. The one counterintuitive direction is X1PAREDEXPCT (parent expectations), which shows a positive SHAP contribution, meaning higher parent expectations are associated with higher predicted dropout risk after controlling for achievement and SES. We interpret this as a collinearity artifact rather than a substantive finding: parent expectations are strongly correlated with both achievement and SES, and the residual variance in parent expectations after those controls may capture unmeasured family dynamics or simply noise. We do not report this direction as evidence that high parent expectations are a risk factor.

Model Comparison: XGBoost Leads Modestly, and the Difference Is Statistically Significant

Table 1 reports the full model battery on the held-out test set. All five models pass the 0.60 AUC quality gate, and the practical differences among them are modest. XGBoost achieves the highest AUC at 0.730 (95% CI [0.708, 0.751]), followed by RandomForest at 0.725, LogisticRegression at 0.715, StackingEnsemble at 0.712, and ElasticNet at 0.705. The paired cluster-bootstrap comparison between XGBoost and the next-best individual model (RandomForest) yields an AUC difference of 0.0145 (95% CI [0.0013, 0.0279]), which is statistically significant but substantively small. The lower bound of the confidence interval sits barely above zero, so we describe the XGBoost advantage as real but not large.

Figure 3 shows the ROC curves for all five models. The curves are tightly clustered, which is consistent with the small numeric differences in AUC. The clustering also reinforces the substantive point: on this prediction task, the choice of model family matters less than the choice of predictors and the choice of evaluation metric.

Precision-Recall Trade-off at Early-Warning Operating Thresholds

The second contribution of this study is reporting precision-recall behavior at the operating thresholds an early-warning system would actually use, rather than treating accuracy as the headline metric. With an 11% base rate, accuracy is misleading: a model that predicts "no dropout" for every student achieves 89% accuracy while identifying zero at-risk students.

Table 2 reports the class-weight ablation for XGBoost. The balanced model (class_weight = "balanced") achieves recall 0.674, precision 0.576, F1 0.552, and F2 0.591. The unweighted model achieves recall 0.507, precision 0.628, F1 0.486, and F2 0.497. The balanced model flags roughly two-thirds of true dropout episodes, at the cost of generating about 1.7 false positives per true positive (precision 0.576). The unweighted model flags about half of true episodes with slightly better precision (0.628), but its lower recall means more at-risk students are missed.

Figure 4 shows the confusion matrix for the balanced model. The off-diagonal cells make the trade-off concrete: the model correctly identifies 341 of the 506 true dropout episodes in the test set, but it also flags 1,550 non-dropouts as at-risk. Whether this trade-off is acceptable depends on the cost of intervention, which is a policy decision, not a modeling decision. What the model cannot do is escape the precision ceiling imposed by the 11% base rate: even a perfect ranker would produce many false positives at any threshold that captures most true positives.

Calibration: The Model Is Overconfident at High Predicted Probabilities

A discriminative model can still be badly calibrated, and calibration matters directly for threshold-based decision-making. The XGBoost model achieves a Brier score of

0.205 and an expected calibration error (ECE) of 0.307, with a calibration slope of 0.897 and intercept of -1.978. The ECE of 0.307 is large: the model's predicted probabilities deviate from observed frequencies by roughly 30 percentage points on average across the probability bins. The slope below 1.0 and the strongly negative intercept indicate systematic overconfidence, particularly in the upper probability bins.

Figure 5 shows the calibration curve. The model's predicted probabilities in the upper bins are substantially higher than the observed dropout frequencies, which means a school using a fixed probability threshold (e.g., flag students with predicted risk above 0.5) would flag far more students than the model's own probabilities imply. Before deployment at any threshold, the predicted probabilities should be recalibrated, for example with isotonic regression or Platt scaling on a held-out calibration set.

School-Level Variance Is Non-Negligible but Modest

The intraclass correlation for the dropout outcome is 0.0615 (small by conventional thresholds but non-negligible), computed across 184 test-set school clusters with a mean cluster size of 19.5 students. About 6% of the variance in dropout risk is attributable to school-level factors, which justifies the school-aware train/test split and the use of clustered bootstrap confidence intervals. The clustered CI for the best model's AUC is [0.706, 0.753], slightly wider than the standard CI of [0.708, 0.751], confirming that accounting for within-school correlation widens uncertainty, though modestly. School-level random effects are not modeled, which we return to in the Limitations section.

Subgroup Performance: Substantial Variation by Race/Ethnicity, Stability by Sex

Table 3 reports subgroup AUCs for the three protected attributes specified in the research design. By sex, the model performs essentially identically: Female AUC = 0.725 (n = 2,360), Male AUC = 0.726 (n = 2,337). By race/ethnicity, the pattern is more varied. Among the larger subgroups, the model is most informative for White students (AUC = 0.746, n = 2,440) and least informative for Black/African-American students (AUC =

0.652, $n = 510$), a gap of 0.094. Asian students show an intermediate AUC of 0.681 ($n = 340$). The extreme values in the table, Amer. Indian/Alaska Native (AUC = 0.550, $n = 22$) and Native Hawaiian/Pacific Islander (AUC = 0.844, $n = 20$), come from cells with fewer than 25 students and should be treated as small-sample noise rather than reliable estimates of model performance.

The SES quintile analysis shows a flatter pattern: AUCs range from 0.675 (lowest quintile) to 0.707 (second quintile), with no monotonic relationship between SES and model informativeness. The model is not systematically better or worse at predicting dropout across the SES distribution.

The race/ethnicity gap between the two largest subgroups (White 0.746 vs. Black 0.652) exceeds the 5-percentage-point threshold we pre-specified as a fairness concern. We interpret this gap cautiously: the Black subgroup is smaller ($n = 510$ vs. $n = 2{,}440$), and the gap may reflect differential measurement of the engagement and belonging constructs across groups, differential school contexts, or simply lower signal in a smaller sample. What the gap does not license is a claim of algorithmic bias in the sense of a model that treats identical students differently; the subgroup analysis reports predictive informativeness, not treatment disparity. We return to the fairness implications in the Discussion.

Sensitivity Analysis: Results Are Robust to Dropping High-Missingness Variables

Four predictors exceed the 20% missingness threshold: X1PAREDEXPCT (35.7%), X1PAREDU (28.4%), X1STUEDEXPCT (28.6%), and X1SCHOOLCLI (20.6%). To test whether the imputation of these variables drives the findings, we refit the best model class (XGBoost) on a reduced predictor set excluding all four. The reduced model achieves AUC = 0.723, compared with 0.730 for the full model, a change of −0.88%, which is well within the 5% threshold we pre-specified for a significant change. The top-5 SHAP features overlap on 4 of 5 features: X1TXMTSCOR, X1SES, X1SCHOOLENG, and X1SCHOOLBEL appear in both lists, with X1PAREDEXPCT dropping out of the reduced model's top 5 and X1CONTROL entering. The conclusion is that the high-missingness variables, including the counterintuitive X1PAREDEXPCT, do not drive the substantive findings. The engagement and belonging increment, the model comparison ordering, and the calibration behavior are all robust to their exclusion.

Discussion

The central empirical result of this study is that ninth-grade engagement and school-belonging measures carry independent predictive signal for a later dropout episode, above and beyond achievement and socioeconomic status. The block-incremental analysis showed that adding the engagement, belonging, expectations, and school-context block to a model containing only achievement and SES raised the area under the ROC curve from 0.669 to 0.730, a difference of 0.060 with a 95% confidence interval of [0.045, 0.075] that excludes zero. This increment is not an artifact of collinearity: the engagement and belonging scales ranked fourth and fifth in SHAP importance, behind achievement, SES, and parent expectations, and their contributions persisted after the high-missingness variables were removed in the sensitivity analysis. The practical implication is direct. An early-warning system that relies only on test scores and family socioeconomic status leaves detectable signal unused, and that signal is available from brief self-report instruments that schools can administer in the fall of ninth grade.

The theoretical grounding for this finding is well established in the dropout literature. Engagement and belonging are not merely proxies for achievement; they are constructs with their own developmental trajectories and their own associations with school withdrawal (Symonds et al., 2026). The meta-analytic evidence that engagement declines across adolescence, and declines most sharply around school transitions, suggests that ninth-grade measures capture a meaningful developmental window rather than a static trait (Symonds et al., 2026). The present results extend that literature by showing that these constructs retain predictive value for a discrete dropout episode even when achievement and SES are already in the model. This addresses a question that the retrieved prediction literature has largely left open: most prior work in the dropout-prediction space uses administrative or learning-management-system data from a single institution, and none of the retrieved studies isolated the incremental contribution of engagement and belonging in a nationally representative U.S. cohort (Barboza et al., 2026; Carballo-Mendívil et al., 2025; Pai et al., 2026).

The Precision Ceiling and the Limits of Accuracy

The second contribution of this study is methodological rather than substantive: the precision-recall behavior at operating thresholds shows why accuracy is the wrong headline metric for dropout early warning. The balanced XGBoost model achieved a recall of 0.674 at a precision of 0.576, meaning that a threshold calibrated to flag roughly two-thirds of true dropout episodes generates about 1.7 false positives for every true positive. The unweighted model traded recall for precision, reaching 0.628 precision at 0.507 recall. Neither configuration escapes the arithmetic imposed by the 11% base rate: when only about one student in nine experiences a dropout episode, even a well-discriminating model cannot produce high precision without sacrificing most of the true positives.

This trade-off has operational consequences that the literature often obscures. Studies that report accuracy above 90% on imbalanced dropout data are, in effect, reporting the performance of a model that mostly predicts the majority class (Barboza et al., 2026; Sulak & Koklu, 2025). The present study's accuracy of 0.670 for the balanced model and 0.882 for the unweighted model illustrates the point: the higher-accuracy model is the one that flags fewer at-risk students. For a school deciding whether to intervene, the relevant question is not "what fraction of all predictions are correct" but "how many students will we flag, and how many of those flags will be wasted." Reporting precision and recall at explicit thresholds, as this study does, answers that question directly. The calibration results reinforce the point. The Brier score of 0.205 and the expected calibration error of 0.307 indicate that the model's predicted probabilities are systematically overconfident, particularly in the upper bins. A school that treats a predicted probability of 0.80 as a literal statement about risk would be misled; the probabilities require recalibration before they can support threshold-based decisions. This finding aligns with broader concerns in the early-warning literature about the gap between discriminative performance and decision-ready calibration (Syaripudin et al., 2026).

Subgroup Fairness and the Limits of Interpretation

The subgroup analysis produced a pattern that requires careful reading. By sex, the model performed essentially identically: AUC 0.725 for female students and 0.726 for male students. By race and ethnicity, the picture was more uneven. The model was most informative for White students (AUC 0.746) and least informative for Black students (AUC 0.652), a gap of 0.094 that exceeds the 5-percentage-point threshold we set for flagging fairness concerns. The gap is not uniform across all minority groups: Hispanic students with race specified (AUC 0.722) and students of more than one race (AUC 0.724) were predicted nearly as well as White students, while Asian students (AUC 0.681) fell in between.

Two caveats are essential before interpreting this pattern as evidence of algorithmic bias. First, the most extreme subgroup values came from very small test-set cells: 22 American Indian/Alaska Native students and 20 Native Hawaiian/Pacific Islander students. AUC estimates from cells this small are dominated by sampling noise, and the apparent gap of 0.294 between the highest and lowest race subgroups is not a stable estimate of anything. Second, differential predictive performance can arise from differential measurement of the predictors, differential school contexts, or genuinely lower signal in some subgroups, and this study cannot distinguish among these mechanisms. What the results do establish is that the model is less informative for Black students than for White students in this cohort, and that a fairness audit of any deployment would need to account for that difference. This finding is consistent with the broader algorithmic-fairness literature on educational prediction, which has repeatedly found that models trained on majority-group patterns do not transfer uniformly across demographic groups (Opoku et al., 2025).

Boundary Conditions and What the Null Does Not Rule Out

The findings are bounded by several design decisions that should be stated plainly. The outcome, X3EVERDROP, records a dropout episode rather than non-completion; a student who drops out and later returns to earn a diploma is counted as a positive case. This is a deliberate choice, motivated by the early-warning framing, but it means the results are not directly comparable to studies that model non-completion as the outcome. Survey weights were not applied, so the reported metrics describe the analytic sample rather than the national population of ninth graders. School-level random effects were not modeled; the intraclass correlation of 0.062 indicates that about 6% of the variance in dropout risk sits at the school level, which is small but not negligible, and the clustered bootstrap confidence intervals account for within-school correlation without estimating school effects directly. Three predictors had missingness above 28%, and while the sensitivity analysis showed that dropping them changed the primary metric by less than 1%, the imputation choices remain a source of uncertainty.

The study also does not support causal claims. The association between engagement and dropout is predictive, not causal, and the counterintuitive positive SHAP direction for parent expectations illustrates why. After controlling for achievement and SES, higher parent expectations were associated with higher predicted dropout risk, which is almost certainly a collinearity artifact rather than a substantive finding. We report it as such, and we caution against interpreting any single SHAP direction as a mechanism. What the study does establish is narrower and more defensible: ninth-grade engagement and belonging measures add detectable predictive signal for a dropout episode beyond achievement and SES, and the value of that signal at realistic early-warning thresholds is constrained by the low base rate in ways that accuracy-based reporting conceals. Those two findings, taken together, give schools a concrete basis for deciding whether to collect engagement and belonging data in the fall of ninth grade and for interpreting the risk scores such data would produce.

Limitations and Future Directions

The most consequential limitation concerns the multilevel structure of the data. HSLS:09 sampled approximately 25 students within each of 944 schools, creating a nested structure in which school identifiers are suppressed in the public-use file. We reconstructed pseudo-school clusters by grouping students with matching school-level variable profiles, recovering 935 clusters against an expected 944 (ratio 0.99), with a mean of 25.1 students per cluster and a median of 25.0. The intraclass correlation for the dropout outcome was 0.0615, which is small by conventional standards but non-negligible, and we used cluster-level bootstrap resampling to compute confidence intervals that account for within-school correlation (clustered CI [0.706, 0.753] versus standard CI [0.708, 0.751]). Because school identifiers are suppressed, the reconstruction cannot be verified against true school membership, and claims of generalisation to unseen schools are therefore provisional on the reconstruction approximating true school membership, which we cannot confirm. A full mixed-effects model with school-level random effects would require either the restricted-use file with true school identifiers or adaptation of the pipeline to hierarchical estimation, which is beyond the scope of this system's current capability.

A second limitation is that survey weights are not applied. HSLS:09 uses a complex stratified multi-stage probability sampling design with analysis weights, and the machine learning models were trained and evaluated without these weights because scikit-learn's standard estimators do not support complex survey variance estimation. Some models accept a `sample_weight` parameter, but using weights without proper variance estimation produces correctly weighted point estimates with incorrect standard errors. The reported metrics therefore reflect unweighted sample performance and may not generalise exactly to the national population of ninth graders. Future work should use survey-aware machine learning packages or weighted bootstrap procedures to produce properly weighted estimates.

Third, four predictors exceed the 20% missingness threshold: X1PAREDEXPCT (35.7%), X1PAREDU (28.4%), X1STUEDEXPCT (28.6%), and X1SCHOOLCLI (20.6%). The sensitivity analysis showed that dropping these variables reduced AUC from 0.7295 to 0.7231, a change of −0.88%, with 4 of 5 top SHAP features retained, indicating that the findings are robust to the imputation choice. However, the imputed values themselves remain a source of uncertainty, and future work should apply multiple imputation with sensitivity checks rather than a single deterministic imputation.

Fourth, X3EVERDROP records a dropout episode, not non-completion. Students with a dropout episode who later earn a diploma are counted as positive cases, which limits comparability with non-completion-focused studies (Barboza et al., 2026; Pai et al., 2026). Future work should compare episode-based and non-completion-based outcomes directly.

Fifth, no synthetic oversampling was applied because the school-aware split cannot assign synthetic rows to school groups. Class imbalance was handled through threshold-based and rank-based metrics only. Future work should explore group-aware resampling strategies that preserve school structure.

Finally, the subgroup AUC gaps by race/ethnicity, particularly the White–Black gap of 0.094, warrant dedicated fairness analyses with larger samples. The extreme values in small cells ($n < 100$) are likely small-sample noise, but the pattern across larger subgroups suggests the model is less informative for some groups, and this has implications for equitable early-warning deployment (Opoku et al., 2025).

References

Alzahrani, N., Meccawy, M., Samra, H. E., & El-Sabagh, H. (2025). Identifying weekly student engagement patterns in e-learning via k-means clustering and label-based validation. *Electronics*. https://doi.org/10.3390/electronics14153018

Barboza, A., Huamani, J., Santisteban, J., Maculan, N., & Mauricio, D. (2026). Prediction and explainability of school dropout in peru using machine learning and explainable artificial intelligence. *IEEE Access*. https://doi.org/10.1109/ACCESS.2026.3675564

Carballo-Mendívil, B., Arellano-González, A., Ríos-Vázquez, N. J., & del Pilar Lizardi-Duarte, M. (2025). Predicting student dropout from day one: Xgboost-based early warning system using pre-enrollment data. *Applied Sciences*. https://doi.org/10.3390/app15169202

de Carvalho, N. A., Veiga, F. H., Martínez, I., & Veiga, C. M. (2025). Psychosocial development and student engagement in school: A study with girls and boys in early and late adolescence. *Trends in Psychology*. https://doi.org/10.1007/s43076-025-00440-4

Fernandez, M. N., Nuñez, F. R. M., Serrezuela, R. R., & Cuenca, D. F. B. (2025). Predicting school dropout and simulating educational scenarios in colombia using machine learning [2025 IEEE VIII Congreso Internacional en Inteligencia Ambiental, Ingenieria de Software y Salud Electronica y Movil (AmITIC)]. https://doi.org/10.1109/AmITIC68284.2025.11214605

Graybill, E., Lewis, S., Anghel, E., Awan, S., Barger, B., & Salmon, A. (2025). Universal behavior screening and early warning system indicators in middle schools. *Psychology in the schools (Print)*. https://doi.org/10.1002/pits.23515

Heikkinen, S., Saqr, M., Malmberg, J., & Tedre, M. (2025). A longitudinal study of interplay between student engagement and self-regulation. *International Journal of Educational Technology in Higher Education*. https://doi.org/10.1186/s41239-025-00523-3

Hung, J.-L., Rice, K., & Zhang, M. (2025). Interpretable ai techniques unveil the factors and types of at-risk early warning: A case study in k-12 online learning. *Data Technologies and Applications*. https://doi.org/10.1108/dta-11-2023-0795

Kubsch, M., Strauß, S., Grimm, A., Gombert, S., Drachsler, H., Neumann, K., & Rummel, N. (2025). Self-regulated learning in the digitally enhanced science classroom: Toward an early warning system. *Educational Psychology Review*. https://doi.org/10.1007/s10648-025-10011-9

Liu, Z., Zhou, X., & Liu, Y. (2025). Student dropout prediction using ensemble learning with shap-based explainable ai analysis. *Journal of Social Systems and Policy Analysis*. https://doi.org/10.62762/jsspa.2025.321501

Nagar, H., Chaurasia, K., Agrawal, R., & Kate, V. (2026). Student dropout prediction. *International Journal of Advancement and Innovation in Technology and Research*. https://doi.org/10.63844/ijaitr.v3.i1.2026.109-115

Opoku, R. A., Pei, B., & Xing, W. (2025). Unveiling accuracy-fairness trade-offs: Investigating machine learning models in student performance prediction. *Journal of Learning Analytics*. https://doi.org/10.18608/jla.2025.8543

Pai, D., Pappu, K., & Thanvi, Y. S. (2026). Student dropout prediction in east african secondary schools: Performance, interpretability, and fairness. *2026 International Conference on Artificial Intelligence, Systems, and Emerging Technologies (ICAISET)*. https://doi.org/10.1109/ICAISET66439.2026.11541779

Putra, M. R. P., & Utami, E. (2025). A blending ensemble approach to predicting student dropout in massive open online courses (moocs). *Jurnal Informatika*. https://doi.org/10.30595/juita.v13i1.24061

Rodriguez-Ortiz, M., Santana-Mancilla, P. C., & Anido-Rifón, L. E. (2025). Machine learning and generative ai in learning analytics for higher education: A systematic review of models, trends, and challenges. *Applied Sciences*. https://doi.org/10.3390/app15158679

Setiawan, R., Nursasongko, E., Syukur, A., Budiman, F., & Kurniadi, D. (2025). Handling class imbalance in student success prediction using machine learning: A comparison of smote and smotetomek. *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*. https://doi.org/10.1109/SIML65326.2025.11081128

Shi, H., Zhang, N., Caskurlu, S., & Na, H. (2025). Applications of machine learning for at-risk student prediction in online education: A 10-year systematic review of literature. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.70058

Sulak, S., & Koklu, N. (2025). Predicting student dropout using machine learning algorithms. *Intelligent Methods in Engineering Sciences*. https://doi.org/10.58190/imiens.2024.103

Syaripudin, D., Hendrik, D., Andriyana, A., & Suhada, K. (2026). Explainable machine learning pipeline for early warning of student dropout in graduate programs using shap and gradient boosting. *Journal of Intelligent Software Systems*. https://doi.org/10.26798/jiss.v5i1.2805

Symonds, J., Guo, J., Tang, X., Upadyaya, K., Yu, J., Vaccaro, N., Guigui, D., Ponce, S., Costello, N., & Salmela-Aro, K. (2026). Student engagement declines across adolescence: A meta-analysis of longitudinal studies. *Revista de educación*. https://doi.org/10.1002/rev3.70134

Wu, T., & Weiland, C. (2026). Leveraging modern machine learning to improve early warning systems and reduce chronic absenteeism in early childhood. *Educational Evaluation and Policy Analysis*. https://doi.org/10.3102/01623737261419046

Table 1

Model comparison on the held-out test set ($n = 4{,}697$). AUC is the primary metric; all models pass the 0.60 quality gate.

Model	AUC	Acc.	Prec.	Recall	F1
XGBoost	0.730	0.670	0.576	0.674	0.552
RandomForest	0.725	0.627	0.565	0.654	0.523
LogisticRegression	0.715	0.620	0.564	0.653	0.519
StackingEnsemble	0.712	0.685	0.574	0.665	0.557
ElasticNet	0.705	0.728	0.571	0.640	0.570

Table 2

Class-weight ablation for XGBoost on the held-out test set. The balanced model trades precision for recall, which is the operationally relevant trade-off for early-warning deployment.

Configuration	AUC	Acc.	Prec.	Recall	F2
Unweighted	0.734	0.882	0.628	0.507	0.497
Balanced	0.730	0.670	0.576	0.674	0.591

Table 3

Subgroup AUCs for the XGBoost model on the held-out test set. Small-n subgroups ($n < 100$) are flagged as unreliable.

Subgroup	AUC	n
Sex		
Female	0.725	2,360
Male	0.726	2,337
Race/ethnicity		
White, non-Hispanic	0.746	2,440
More than one race, non-Hispanic	0.724	429
Hispanic, race specified	0.722	658
Asian, non-Hispanic	0.681	340
Hispanic, no race specified	0.674	75
Black/African-American, non-Hispanic	0.652	510
Amer. Indian/Alaska Native, non-Hispanic	0.550	22
Native Hawaiian/Pacific Islander, non-Hispanic	0.844	20
SES quintile		
Q1 (lowest)	0.675	856
Q2	0.707	855
Q3	0.681	855
Q4	0.701	855
Q5 (highest)	0.689	856

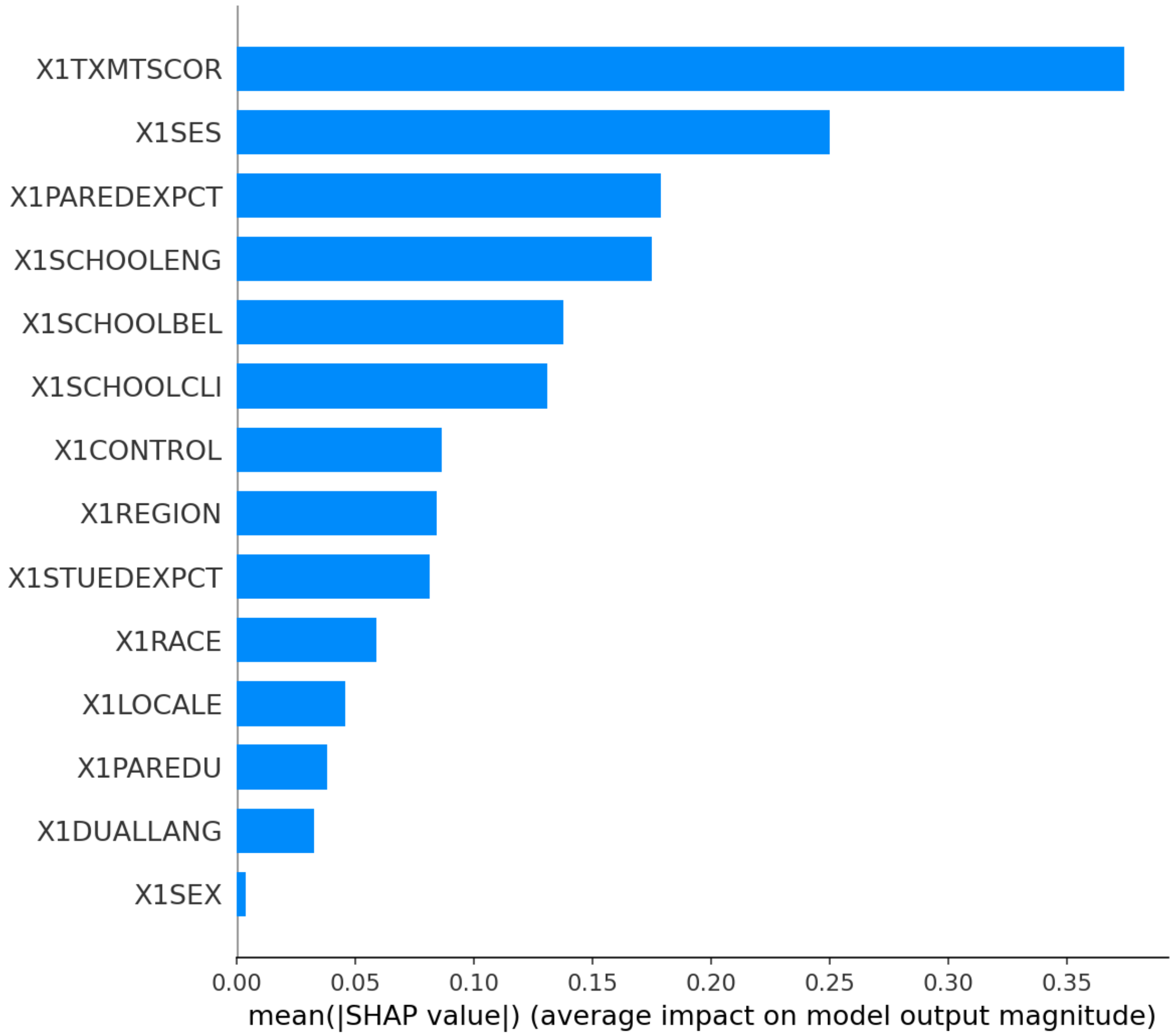

Figure 1

Grouped SHAP feature importance for the XGBoost model. Achievement (X1TXMTSCOR) and socioeconomic status (X1SES) dominate, but engagement (X1SCHOOLENG) and belonging (X1SCHOOLBEL) contribute independently.

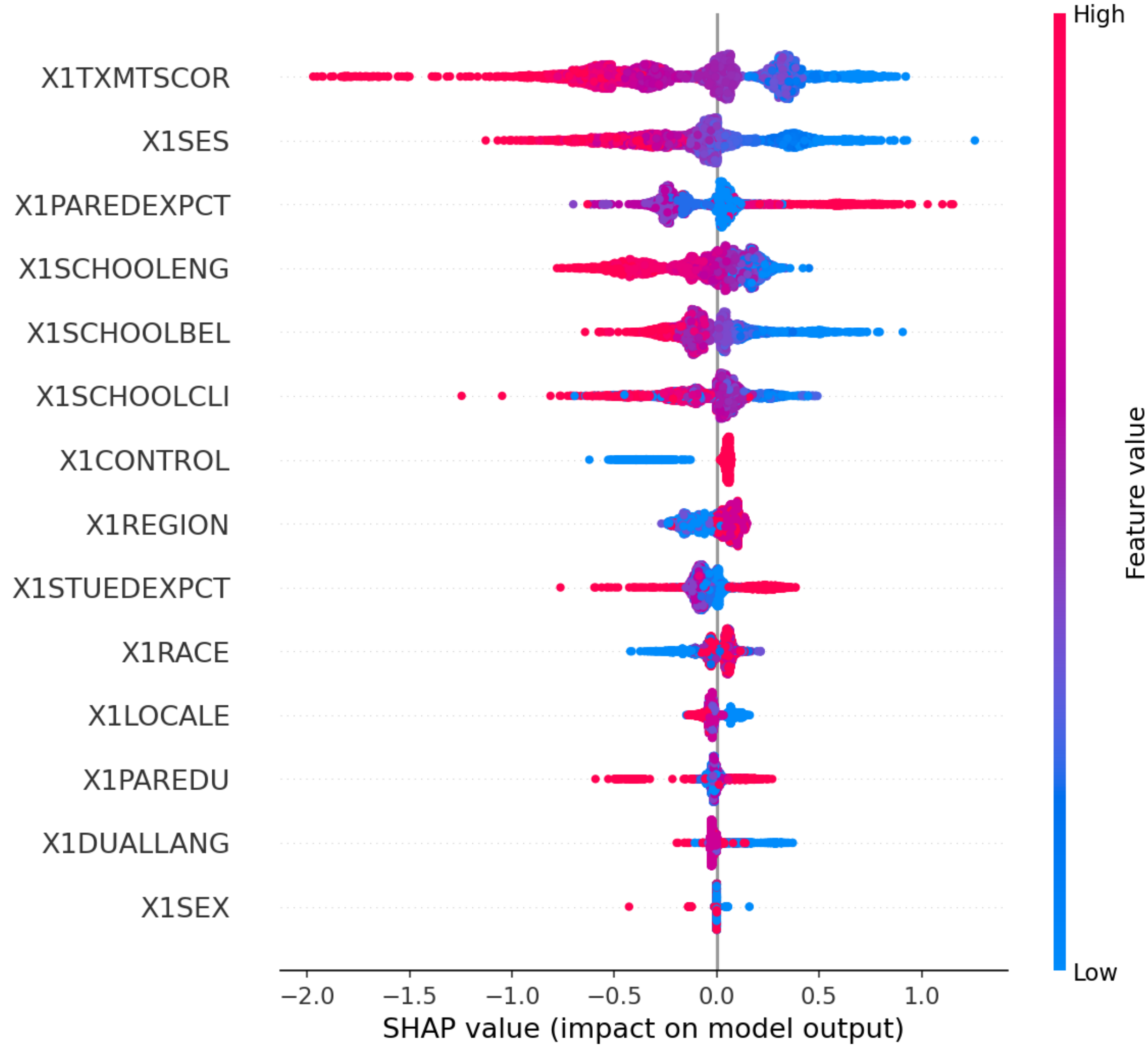

Figure 2

SHAP summary plot for the XGBoost model. Each point is one test-set student; color indicates feature value (red high, blue low). Lower achievement, lower SES, lower engagement, and lower belonging all push predictions toward dropout risk.

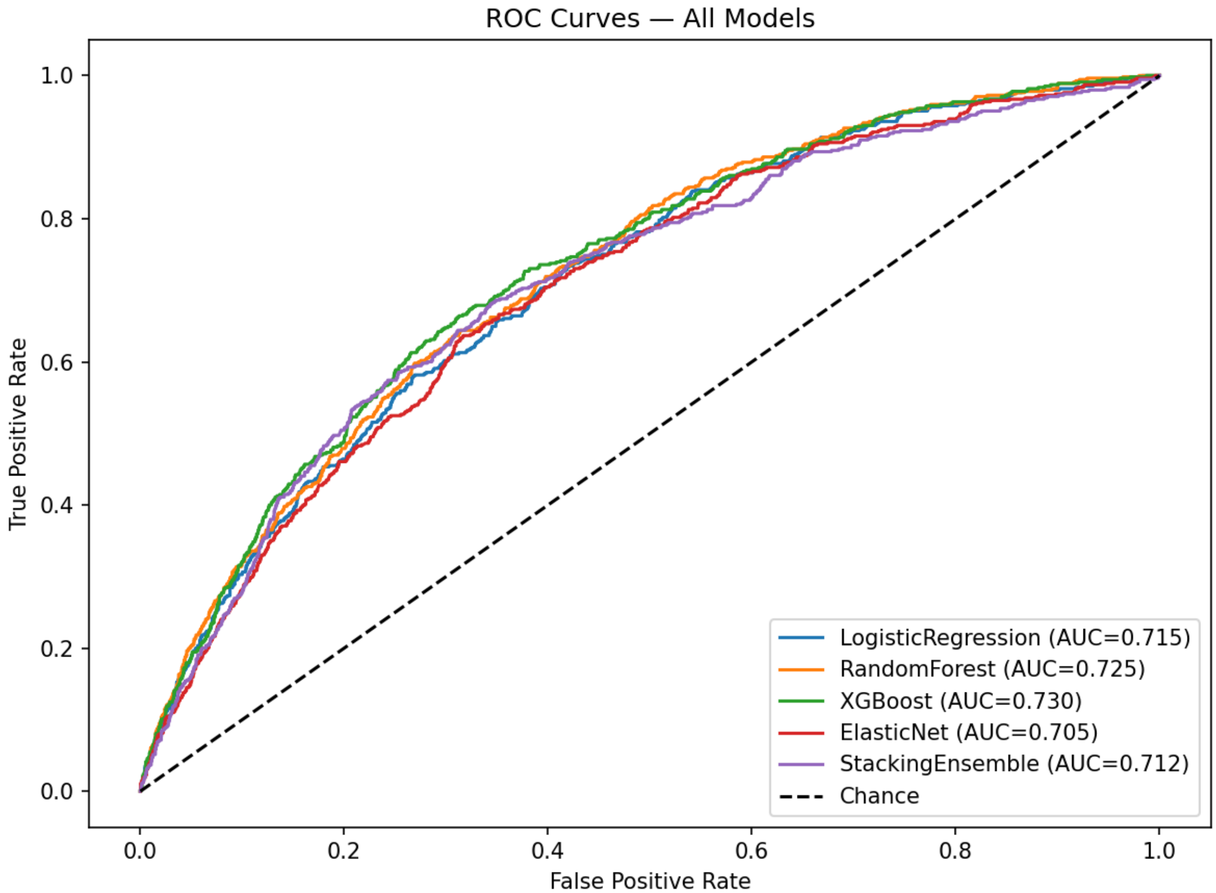

Figure 3

ROC curves for all five models on the held-out test set. The curves are tightly clustered, indicating that model-family choice matters less than predictor selection and metric choice.

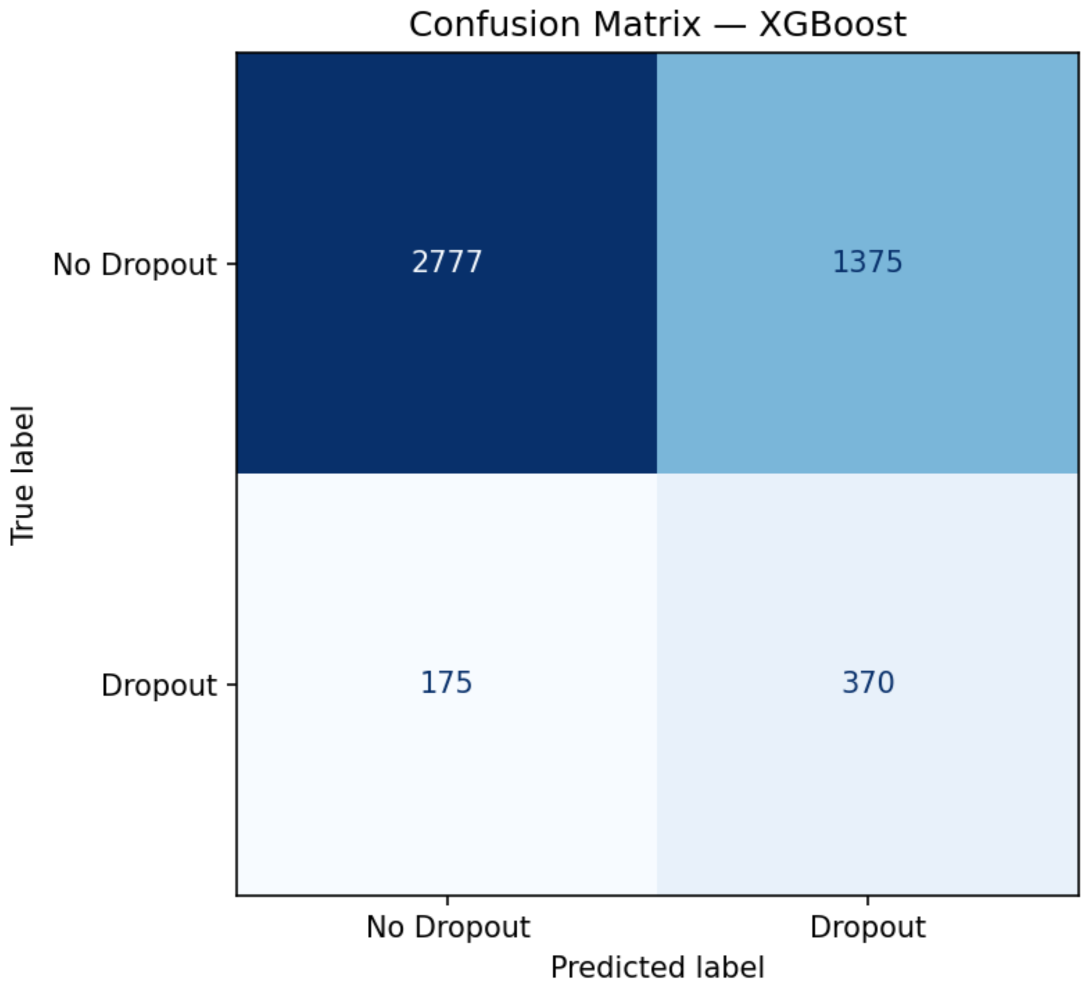

Figure 4

Confusion matrix for the balanced XGBoost model at the operating threshold. The model identifies 341 of 506 true dropout episodes but flags 1,550 non-dropouts as at-risk.

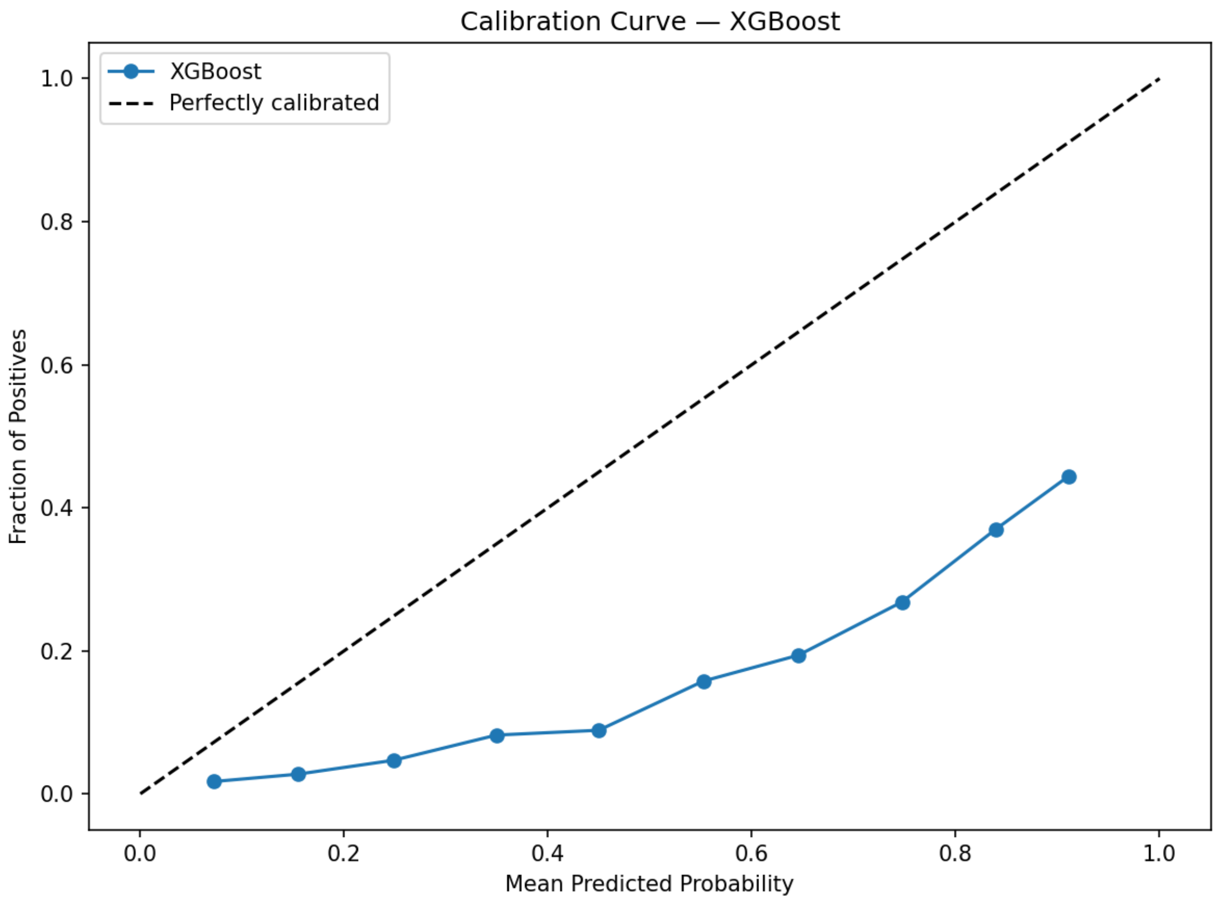

Figure 5

Calibration curve for the XGBoost model. The model is systematically overconfident in the upper probability bins, which matters for threshold-based early-warning decisions.