

# Multimodal Machine Learning for Student Retention Prediction: Integrating Temporal, Textual, and Tabular Features

**Kashaina Nucum**

Tennessee Technological University  
Department of Computer Science  
kashainanucum@gmail.com

## Abstract

Student retention analysis and prediction can support earlier and more targeted intervention in higher education. This study presents a retention analytics framework for first-year undergraduate student retention prediction in the College of Engineering at Tennessee Technological University. The system integrates structured sociodemographic and academic performance features with advisement-note features derived from natural language processing. Tree-based machine-learning models are used to support early-warning prediction, while SHAP-based explainability are applied to identify the factors most strongly influencing individual and population-level retention risk.

## Introduction

Student retention is a major concern in engineering education because students face academic, institutional, and personal barriers during the transition to college. Early identification of at-risk students can support earlier intervention. While many studies rely on demographic and academic records, advisement notes may capture additional barriers such as financial hardship, mental health concerns, family responsibilities, commuting issues, and withdrawal intentions.

This work addresses two questions: which features most strongly influence individual retention predictions, and which factors are most consistently associated with attrition across the student population.

## Related Work

Prior studies have used demographic, academic, and temporal data to predict student dropout, often achieving strong predictive performance. Some work has also incorporated advising notes and deep learning methods such as BERT, BiLSTMs, and LSTMs (Alam, 2021). Other studies focus more narrowly on socio-demographic or course-performance data (Jayaraman, 2021; Trivedi, 2022; Niyogisubizo et al., 2022; Glandorf et al., 2022). This study extends that line of work by combining structured first-year features with advisement-note signals in an interpretable framework.

Copyright © 2026 by the authors. Open access article published under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

## Methodology

The dataset included College of Engineering students enrolled from Fall 2021 and Spring 2025. After cleaning, 2,328 students remained, with a retention rate of 77.1% and dropout rate of 22.9%. The 2024 cohort was reserved for testing.

## Structured Features

The structured pipeline included binary, categorical, and continuous predictors from student background and first-year academic records. Features included residency, Pell eligibility, first-generation status, semester, major, gender, term GPA, credit hours, course count, class standing, pass rate, DFW rate, online course rate, minimum grade, high school GPA, and ACT scores.

The analysis also included selected gateway STEM courses with high DFW rates. Instead of modeling all courses, binary indicators were created for whether a student took a high-risk course and whether they earned a DFW in that course. For longer-term patterns, features from the two most recent semesters were used rather than a full sequence model for simplicity and interpretability.

## Advisement Note Modeling

Because labeled note data were limited, a lightweight multi-label NLP framework was used. The note pipeline focused on nine retention-relevant categories, including academic struggle, mental health, financial crisis, physical illness, external obligations, commute or housing risk, family obligations, social isolation, and withdrawal risk.

Each category was modeled as an independent binary SetFit classifier using manually specified positives and routine advising negatives. Notes were also assigned an overall sentiment score using `cardiffnlp/twitter-roberta-base-sentiment-latest`. Outputs were aggregated to the student-term level as interpretable tabular predictors, including category flags, category counts, and sentiment summaries.

## Predictive Modeling

Multiple models were evaluated, including XGBoost, LightGBM, CatBoost, Random Forest, Extra Trees, and Logis-

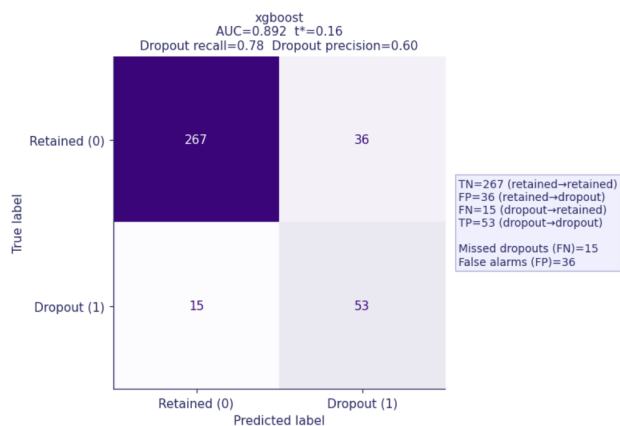


Figure 1: Confusion matrix for the final XGBoost first-term retention model with advisement-note features

tic Regression along with SMOTE if applicable. Since retention prediction emphasizes identifying at-risk students, dropout recall was treated as especially important alongside AUC, precision, balanced accuracy, and Brier score.

## Results

Model comparison showed that several methods remained competitive after advisement-note features were added. However, SMOTE did not consistently improve generalization. In particular, LightGBM + SMOTE and XGBoost + SMOTE achieved very high training AUC values of 0.9594 and 0.9541, respectively, but their test AUC values dropped to 0.8705 and 0.8736, indicating overfitting. By contrast, the stronger non-SMOTE models showed smaller train-to-test gaps and more stable performance.

Among the non-SMOTE candidates, several models performed similarly. Logistic Regression achieved a test AUC of 0.8965 with a dropout recall of 0.7794, Random Forest achieved a test AUC of 0.8936 with a recall of 0.7941, and Extra Trees achieved a test AUC of 0.8940 with a recall of 0.7794. XGBoost remained highly competitive, with a test AUC of 0.8957 and a dropout recall of 0.8088, offering a stronger balance between discrimination and sensitivity to at-risk students than many alternatives.

XGBoost was selected as the final model because it combined strong discrimination, strong dropout detection, and more stable generalization than the SMOTE-based boosted models. Its confusion matrix is shown in Figure 1. On the test set, the model achieved an AUC of 0.8957 and a dropout recall of 0.8088, showing that it was able to distinguish retained from withdrawn students well while still identifying a large share of students who ultimately dropped out. Its test precision was 59.6%, indicating a moderate false-positive rate, although some of those flagged students may still benefit from intervention.

## Explainability

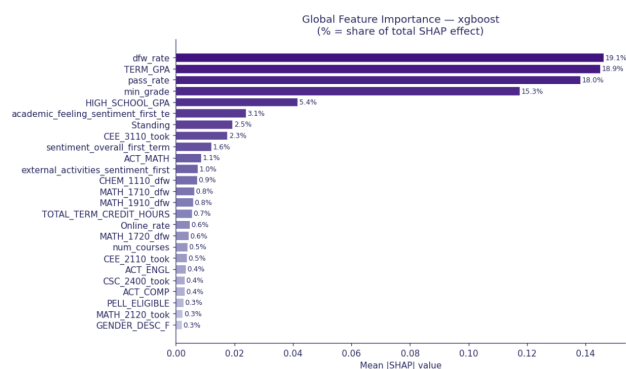


Figure 2: SHAP summary for the XGBoost model with advisement-note features

SHAP results showed that early academic performance variables were the strongest drivers of first-term withdrawal prediction. DFW rate, term GPA, pass rate, and minimum grade were the most influential features, while high school GPA, a pre-college variable, had a smaller effect. Among note-derived variables, `academic_feeling_sentiment_first_term` was one of the more influential predictors, especially in the negative direction. These results suggest that note features add useful context, but structured academic measures remain the dominant predictive signal.

## Future Work

Longitudinal modeling components remain under development. Additionally, while this study adopted several simplified design choices to maintain a practical and interpretable framework, longer-term future work could address additional edge cases and expand the feature space, such as double majors, age, non-traditional student flag, and students enrolled in pre-engineering coursework, in meaningful ways.

## Conclusion

This work presents a retention prediction framework that combines structured academic and demographic data with advisement-note signals. Results indicate that first-year academic performance, especially pass rate, DFW rate, term GPA, and minimum grade, is more strongly associated with dropout than most demographic characteristics. Advisement-note features added useful contextual information, while XGBoost showed promising held-out performance.

## Acknowledgments

The authors acknowledge the Research Computing and Data division at Tennessee Tech University (RRID:SCR027555) for providing computational resources and support services that have contributed to the research results reported within this paper. Specifically, computations were supported by the National Science Foundation under Award #2127188.

## References

- Xu, J.; Moon, K. H.; and van der Schaar, M. 2017. A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, 11(5):742–753.
- Prabowo, H.; Hidayat, A. A.; Cenggoro, T. W.; Rahutomo, R.; Purwandari, K.; and Pardamean, B. 2021. Aggregating time series and tabular data in deep learning model for university students' GPA prediction. *IEEE Access*, 9:87370–87377.
- Niyogisubizo, J.; Liao, L.; Nziyumva, E.; Murwanashyaka, E.; and Nshimyumukiza, P. C. 2022. Predicting students' dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, 3:100066.
- Jayaraman, J. D. 2020. Predicting student dropout by mining advisor notes. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, 629–632.
- Alam, M. A. U. 2021. College student retention risk analysis from educational database using multi-task multi-modal neural fusion. *arXiv preprint arXiv:2109.05178*.
- Glandorf, D.; Lee, H. R.; Avakian-Orona, G.; Pumptow, M.; Yu, R.; and Fischer, C. 2024. Temporal and between-group variability in college dropout prediction. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24)*, 486–497. ACM.