Proceedings of CONF-CDS 2026 Symposium: Machine Learning and Neural Network Applications in Engineering
DOI: 10.54254/2755-2721/2026.AST37155

Predicting University Student Dropout Using Logistic Regression: Comparing Enrollment and Semester-Based Models

Yanping Feng

Faculty of Science, University of New South Wales, Sydney, Australia
N2601324H@e.ntu.edu.sg

Abstract. A substantial body of academic research has examined the determinants of undergraduate attrition and premature program departure. This study investigates whether logistic regression could identify students who discontinued their programs compared to those who graduated. An 80/20 stratified train/test split was used to evaluate three logistic regression models based upon 3,630 resolved outcomes, where 1,421 were designated as Dropout and 2,209 as Graduate. Model E included all enrolment variables. First-semester academic records were added to Model S1. Second-semester academic records were added to Model S2. The receiver operating characteristic curve (ROC-AUC) values were 0.836, 0.917 and 0.935, respectively. The Brier scores were 0.155, 0.099 and 0.080, respectively. Older age, male gender, and debtor status were positively related to dropout odds. As exact dropout timestamps are unavailable, semester-based models are treated as post facto. These results support targeted assistance to students experiencing difficulty rather than inferring causality.

Keywords: student dropout, logistic regression, higher education, predictive modeling, academic performance

1. Introduction

The loss of a student can result in consequences for that student and the university. Research identified academic preparedness, economic pressure, academic progress, and personal circumstances as contributing factors to leaving college. Few studies directly compared enrolment data with subsequent semester data within an interpretable model. A predictive model should only be utilized to support a human decision about providing assistance to students.

This study uses the Predict Students' Dropout and Academic Success dataset [1]. It explores whether logistic regression can distinguish dropout from graduation and analyzes how much first- and second-semester academic results improve prediction over enrollment information. Three nested models are compared using discrimination, probability accuracy, calibration, and threshold-based classification. This comparison shows whether improved accuracy is available early enough to support intervention.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AST37155

2. Literature review

Although Tinto's model does not yield a predictive formula, it provides a theoretical justification for separating enrollment variables from semester results. Students enter university with different levels of preparation, goals, and personal conditions. Their later academic and social integration can reinforce or undermine commitment; semester grades therefore reflect integration after study begins [2].

Student dropout is associated with academic, economic, institutional, and personal factors. Socio-economic resources and the costs and benefits of university study influence dropout decisions [3]. Previous academic achievement, study conditions, institutional factors, and personal characteristics also appear consistently in prior studies [4]. These findings support including admission grades, age, scholarship status, debt status, tuition payment status, and course in the models.

Prior work in the same institutional setting examined early prediction of academic success [5]. Machine learning studies also show that dropout risk can be predicted and stress the practical value of early identification [6]. Retention models should provide actionable information for educational institutions [7].

Logistic regression provides predicted probabilities and interpretable odds ratios for the relationship between predictors and dropout. This makes it suitable for comparing model performance while explaining the direction of associations [8].

Model performance should be assessed using multiple metrics. ROC-AUC measures discrimination between dropout and graduation, while PR-AUC evaluates precision for the positive class [9, 10]. The Brier score measures probability error, and calibration intercepts and slopes show whether predicted probabilities are systematically inaccurate [11].

3. Methodology

3.1. Data source and analytic sample

Data were obtained from the UCI Machine Learning Repository [1]. The dataset contained 4,424 rows and 37 columns, with no missing values or duplicated rows. Outcomes were Graduate (n = 2,209), Dropout (n = 1,421), and Enrolled (n = 794). Enrolled cases were removed because their final status was unresolved. The final sample contained 3,630 students, with Graduate coded 0 and Dropout coded 1. The resolved-case dropout rate was 39.1%.

3.2. Predictors and model specification

Model E included age at enrollment and admission, previous qualification grades, application order, course, gender, scholarship, debtor, and tuition payment status. Model S1 added first-semester approved units and average grade, while Model S2 added the corresponding second-semester measures. Attendance type was excluded because it was perfectly collinear with course. Nursing was the largest program and served as the reference category.

3.3. Preprocessing and logistic regression

The sample was split into 80% training data and 20% test data using stratified sampling and random seed 1100, providing 2,904 training and 726 test observations. Preprocessing was fitted within the training data. Continuous predictors were median-imputed and standardized, while categorical

DOI: 10.54254/2755-2721/2026.AST37155

predictors were mode-imputed and converted to indicator variables. Logistic regression estimated dropout probability using Equation (1).

$$\log[p/(1-p)] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \quad (1)$$

3.4. Validation and threshold selection

Repeated stratified fivefold cross-validation was conducted on the training set and repeated five times. Thresholds from 0.10 to 0.90 were examined using out-of-fold training predictions. The selected thresholds were 0.28 for Model E, 0.36 for Model S1, and 0.42 for Model S2. Each pipeline was then fitted to the full training set and evaluated on the untouched test set.

3.5. Model evaluation and interpretation

Final evaluation used discrimination, calibration, and classification metrics. Odds ratios and 95% confidence intervals were reported for selected Model S1 predictors. Model S1 was used because correlated academic measures in Model S2 made individual semester effects harder to interpret. All Model S1 variance inflation factors were below 2.6.

3.6. Ethics and limitations

Model S2 recall and false-positive rates were compared by gender and age group. This dataset reflects one university and does not include demographic characteristics. The data cannot establish casual relationships, nor did the dropout date exist. Therefore, semester grades may have been recorded after students had exited; Models S1 and S2 were treated as retrospective classification models.

4. Results

There were 2,209 graduates and 1,421 dropouts in the sample. Dropouts were older, had completed fewer units for approval by graduation, and earned lower semester grades. A total of 67.8% of dropouts versus 98.7% of graduates were paying tuition at the same time. Debtors represented 22% of dropouts versus 5% of graduates. The descriptive characteristics of the resolved-outcome analytic sample are displayed in Table 1.

Table 1. Characteristics of the resolved-outcome analytic sample

Characteristic	Overall	Graduate	Dropout
Students, n	3,630	2,209	1,421
Age at enrollment, M (SD)	23.46 (7.83)	21.78 (6.69)	26.07 (8.70)
Admission grade, M (SD)	127.29 (14.61)	128.79 (14.07)	124.96 (15.13)
Previous qualification grade, M (SD)	132.92 (13.24)	134.08 (13.34)	131.11 (12.87)
First-semester approved units, M (SD)	4.79 (3.24)	6.23 (2.58)	2.55 (2.86)
First-semester grade, M (SD)	10.53 (5.06)	12.64 (2.70)	7.26 (6.03)
Second-semester approved units, M (SD)	4.52 (3.16)	6.18 (2.27)	1.94 (2.57)
Second-semester grade, M (SD)	10.04 (5.48)	12.70 (2.69)	5.90 (6.12)

DOI: 10.54254/2755-2721/2026.AST37155

Table 1. (continued)

Scholarship holder, n (%)	969 (26.7)	835 (37.8)	134 (9.4)
Debtor, n (%)	413 (11.4)	101 (4.6)	312 (22.0)
Tuition fees up to date, n (%)	3,144 (86.6)	2,180 (98.7)	964 (67.8)
Male, n (%)	1,249 (34.4)	548 (24.8)	701 (49.3)

Note. M = mean; SD = standard deviation. Percentages are within outcome columns; Enrolled cases were excluded.

Validation ROC-AUC increased from 0.843 for Model E to 0.919 for Model S1 and 0.936 for Model S2. Brier scores decreased from 0.153 to 0.096 and 0.081. Table 2 presents the test-set performance of the three models.

Table 2. Test-set performance of the three logistic regression models

Panel A: Discrimination and Calibration					
Model	ROC-AUC	PR-AUC	Brier	Cal. intercept	Cal. slope
Model E	0.836	0.804	0.155	-0.059	0.946
Model S1	0.917	0.914	0.099	-0.013	0.984
Model S2	0.935	0.938	0.080	-0.063	0.976
Panel B: Classification at Training-Selected Thresholds					
Model	Threshold	Accuracy	Precision	Recall	F1
Model E	0.28	0.714	0.595	0.842	0.697
Model S1	0.36	0.861	0.808	0.845	0.826
Model S2	0.42	0.897	0.877	0.856	0.866

Note. ROC-AUC = area under the receiver operating characteristic curve; PR-AUC = average precision; Cal. = calibration. Thresholds were selected from training predictions.

On the test set, Model E achieved ROC-AUC = 0.836. At its 0.28 threshold, recall was 0.842 but precision was 0.595, indicating more false alarms. Model S1 improved ROC-AUC to 0.917 and F1 to 0.826. Model S2 performed best, with ROC-AUC = 0.935, Brier = 0.080, and F1 = 0.866.

Figure 1 displays forest-plot results of adjusted odds ratios from Model S1. In Model S1, older age (OR = 1.53, 95% CI [1.34, 1.76]), male gender (OR = 1.89, 95% CI [1.46, 2.45]), and debtor status (OR = 2.83, 95% CI [1.90, 4.23]) were associated with higher dropout odds. Scholarship status (OR = 0.34, 95% CI [0.25, 0.46]) and current tuition payment status (OR = 0.05, 95% CI [0.03, 0.08]) were associated with lower odds. Approved units and grades were also negatively associated with dropout.

DOI: 10.54254/2755-2721/2026.AST37155

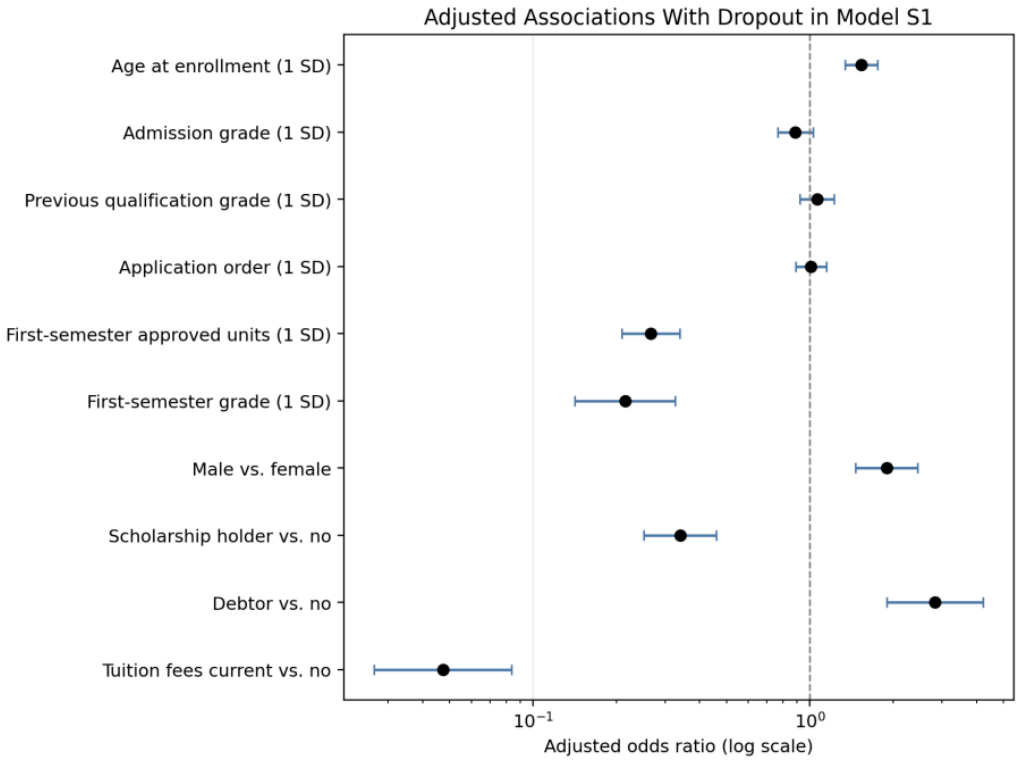

Figure 1. Adjusted odds ratios for selected Model S1 predictors

Note. Points show adjusted odds ratios and 95% confidence intervals. Continuous predictors represent a one-standard-deviation increase; course indicators are omitted.

Model S2 recall was 90.9% for males and 80.9% for females, with false-positive rates of 0.198 and 0.039. Recall was 93.1% for students over 25 and 79.1% for younger students, with false-positive rates of 0.161 and 0.058.

5. Discussion

Adding first-semester data increased ROC-AUC by 0.081 and reduced the Brier score by 0.056. Second-semester data produced further improvements of 0.018 and 0.019. These results are consistent with studies that use academic performance in dropout prediction [5, 6]. However, the most accurate model is not always the best early intervention model because later data delay support.

Model E's ROC-AUC of 0.836 shows that enrollment, financial, demographic, and course variables have predictive value. Using a threshold lower than 0.60 resulted in better recall and poorer precision. Thus, depending on available resources and the relative cost of missing a dropout versus generating a false alarm, threshold selection must occur.

First-semester completion of units, receipt of scholarships, and payment of tuition when due were negatively associated with dropout odds. These findings corresponded with academic integration and socio-economic status [2, 3]. These findings do not provide evidence of causality. Tuition and scholarship status may represent eligibility considerations.

Although the sample with known outcome differed from the population at risk, the data represented one Portuguese institution; thus, generalisability to other environments was limited. Additionally, since exact timing of dropout was unknown, semester grades may document the

DOI: 10.54254/2755-2721/2026.AST37155

leaving process instead of predicting it prospectively. Finally, additional external and temporal validation are necessary prior to application.

Recall was greater for female students and younger students; however, false-positive rates were greater for male students and older students. Universities must monitor calibration, maintain human oversight, and utilise output to offer academic or financial support, not as criteria for limiting enrolment or awarding scholarships [7].

6. Conclusion

Regardless of the inclusion of semester results, the test ROC-AUC score increased from 0.836 for Model E to 0.917 for Model S1 and 0.935 for Model S2. Older age, male gender, and debtor status were positively associated with dropout odds; whereas scholarships and tuition payments were negatively associated. These results support assistance rather than implying causative relationships or restricting access to education. The dataset does not contain information regarding when semester-level results were documented relative to when each student dropped out. Subsequent research should incorporate time-stamped data documenting when students drop out. Additionally, future studies could assess predictive ability at enrollment, post-assessment period 1, and the end of the first semester. This would enable researchers to determine whether earlier predictions could yield more accurate assessments without waiting too long to initiate interventions.

References

[1] Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict students' dropout and academic success [Data set]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C5MC89

[2] Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*(1), 89–125. https://doi.org/10.3102/00346543045001089

[3] Aina, C., Baici, E., Casalone, G., & Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences, 79*, 101102. https://doi.org/10.1016/j.seps.2021.101102

[4] Behr, A., Giese, M., Teguim Kamdjou, H. D., & Theune, K. (2020). Dropping out of university: A literature review. *Review of Education, 8*(2), 614–652. https://doi.org/10.1002/rev3.3202

[5] Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M. T., & Realinho, V. (2021). Early prediction of student's performance in higher education: A case study. In Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, & A. M. R. Correia (Eds.), *Trends and applications in information systems and technologies (Vol. 1*, pp. 166–175). Springer. https://doi.org/10.1007/978-3-030-72657-7_16

[6] Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education, 10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

[7] Addison, L., & Williams, D. (2023). Predicting student retention in higher education institutions (HEIs). *Higher Education, Skills and Work-Based Learning, 13*(5), 865–885. https://doi.org/10.1108/HESWBL-12-2022-0257

[8] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

[9] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[10] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[11] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gönen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology, 21*(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2