technologies

MDPI

Article

Academic Dropout Prediction Using Large-Scale Static Institutional Data: A Multi-Scenario Machine Learning Study

Rômulo Barreto Mincache, Leonardo Gabiato Catharin *, Lucas de Oliveira Teixeira, Thelma Elita Colanzi, Yandre Maldonado e Gomes da Costa and Valéria Delisandra Feltrim

Department of Informatics, State University of Maringá (UEM), Maringá 87020-900, PR, Brazil; romulo.mincache@gmail.com (R.B.M.); loteixeira@uem.br (L.d.O.T.); teclopes@uem.br (T.E.C.); ymgcosta@uem.br (Y.M.e.G.d.C.); vdfeltrim@uem.br (V.D.F.)
* Correspondence: pg55751@uem.br

Abstract

Academic dropout is costly for students and institutions, and support actions are most effective early, when little information beyond enrollment records is available. However, the most informative predictors are derived from academic trajectory data, such as grades and course progression, which only become available after students have completed one or more terms. This study investigates dropout prediction using supervised machine learning (ML) applied to records of 66,820 students across 45 undergraduate programs of the State University of Maringá, Brazil (2002–2022), trained exclusively on static enrollment-time variables, since these were the only institutional data available. Although this restricts the information the models can use, it also allows students who may be at risk to be flagged very early. Decision Tree, Random Forest, and eXtreme Gradient Boosting (XGBoost) models were evaluated in ten training configurations, covering the complete dataset, the complete-generations subset, a temporal cohort split, academic centers, individual programs, and resampled variants. Evaluation was based on per-class precision, recall, and F1-score, together with threshold-free and calibration measures (PR-AUC, ROC-AUC, and Brier score). XGBoost performed best in the institution-wide configurations, reaching a dropout recall of 0.64 at a precision of 0.51 with undersampling and, without resampling, a PR-AUC of 0.586 against a dropout prevalence of 0.364, which corresponds to 1.61 times the performance of random ranking. Under the temporal split, recall at the default threshold fell from 0.37 to 0.18, while ROC-AUC changed little (0.708 to 0.679); this loss reflects miscalibration under changing dropout prevalence and can be corrected without retraining the model. The main contributions are an evaluation methodology for enrollment-time dropout prediction that controls information leakage and includes temporal validation, and a deployment protocol that uses the model outputs to prioritize student support when follow-up capacity is limited.

Keywords: academic dropout; machine learning; classification; institutional data; static features; class imbalance; educational data mining

Academic Editor: Pedro Antonio Gutiérrez

Received: 13 June 2026
Revised: 28 August 2026
Accepted: 1 September 2026
Published: 7 September 2026

1. Introduction

Academic dropout is a persistent problem in higher education, with consequences for students, institutions, and society [1]. When students leave their programs before completion, institutions lose part of the resources invested in each admission, and students may face setbacks in their academic and professional lives [2]. Anticipating dropout early enough to act on it has become an important research problem.

Dropout is a multifactorial phenomenon, influenced by socioeconomic, academic, psychological, and institutional factors [3–5]. Several theoretical models help explain how these factors affect dropout after enrollment. Tinto's student integration model [3], for example, relates persistence to students' academic and social integration into the institution, and Bandura's social cognitive theory [6] emphasizes factors such as self-efficacy and the interaction between individuals and their environment. The enrollment-time variables used in this study characterize students' initial demographic, educational, and institutional context, but do not directly measure academic and social integration, self-efficacy, or other processes that develop after enrollment.

Dropout can also be analyzed at different levels, including programs, institutions, and higher education systems [7], and may be characterized as temporary or permanent [8]. In addition, dropout patterns may differ between face-to-face and distance education because of differences in student engagement, access to institutional resources, and academic support [9–12].

Machine learning (ML) techniques have been widely applied to educational data for dropout prediction. Previous studies have evaluated decision trees, Random Forest, eXtreme Gradient Boosting (XGBoost), Bayesian networks, logistic regression, neural networks, and other methods [13–18]. Tree-based models are common in studies based on institutional data because they handle structured tabular data and nonlinear relationships effectively [14,19]. However, no single algorithm performs best across all educational settings [13,15], so several models are compared here under the same evaluation protocol.

Dropout prediction models may use both static and dynamic data. Static variables include demographic characteristics and information available at admission, whereas dynamic variables include grades, attendance, and academic progression [4,20]. Dynamic data provide information about students' trajectories and often improve predictive performance, but they only become available after students have begun their programs and may be inconsistently recorded in institutional databases. Enrollment-time data therefore remain relevant when the goal is to identify students who may need support as early as possible [19].

Despite recent advances, three limitations motivate the present study. First, many studies use random train-test splits [14,15], which do not reproduce a real-world setting [21,22]. Evaluations restricted to classification metrics at a fixed decision threshold also provide little information about model discrimination and calibration. Second, although institution-wide models are common, differences across academic centers and programs have received less attention in large-scale studies [16]. Third, there is little evidence on how models restricted to enrollment-time data behave under temporal and institutional variation.

To address these limitations, we investigate these questions using static institutional data from the State University of Maringá (UEM), Brazil, covering a 20-year period and more than 66,000 students. Multiple experimental configurations are compared to examine how temporal variation, institutional context, and class imbalance affect predictive performance.

The main contributions of this study are as follows:

- A systematic comparison of Decision Tree, Random Forest, and XGBoost models across institution-wide, temporal, academic center, and program-level configurations using only enrollment-time data.
- An analysis of class imbalance strategies using both threshold-dependent and threshold-free evaluation measures, showing how resampling affects model predictions.
- An analysis of feature importance using tree-based importance measures and SHAP to examine which enrollment-time variables contribute most to model predictions and how their attributed importance changes across cohorts.

- A temporal evaluation using both a fixed cohort split and rolling validation to assess model performance on future student cohorts and its implications for model deployment.
- A fairness and institutional-bias analysis evaluating whether predictive errors differ systematically across demographic and admission-related student groups.

2. Materials and Methods

This section describes the data preparation and analysis procedures, cohort definition, experimental design, ML models, performance evaluation, and model interpretation methods used in this study. Figure 1 provides an overview of the main steps of the experimental pipeline.

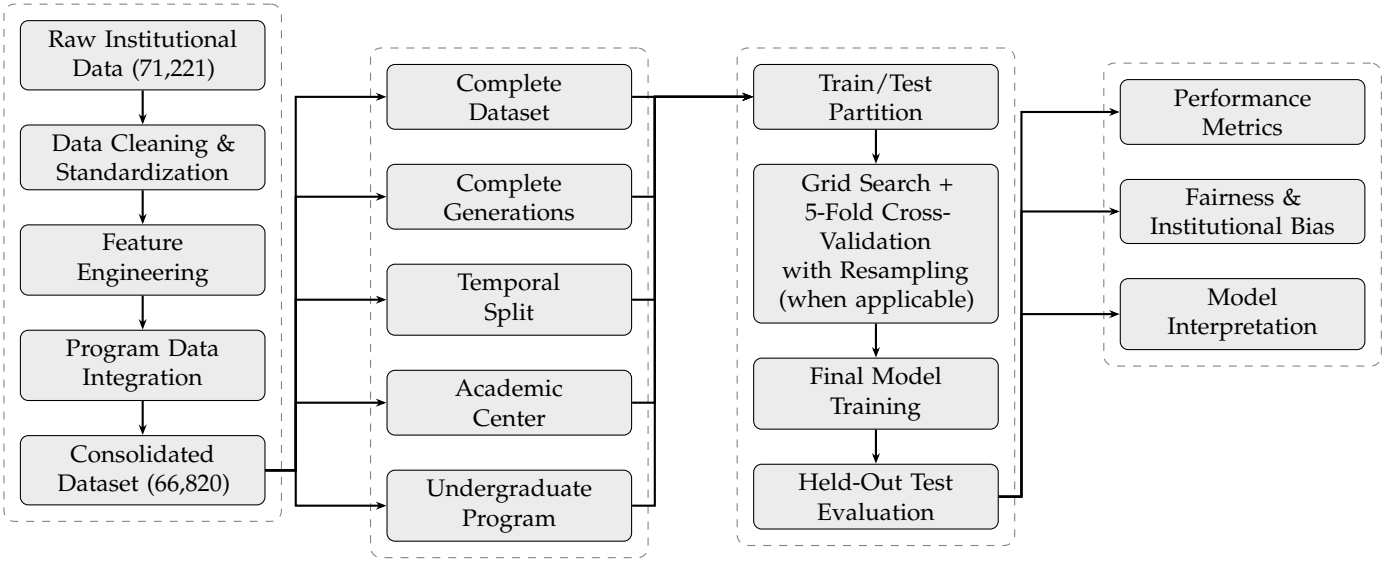


Figure 1. Overview of the experimental pipeline. The dashed line boxes group the sequential stages of the methodological pipeline (Data Preprocessing, Dataset Splitting, Model Training & Validation, and Output Evaluation and Interpretation).

2.1. Dataset

The dataset used in this study was provided by the Data Processing Center (DPC) of UEM and was compiled by integrating data from multiple internal institutional databases. Before being provided to the researchers, all records were anonymized by the DPC, with personally identifiable information removed in accordance with applicable data protection regulations.

The original dataset contains records from 71,221 students enrolled in on-campus undergraduate programs between 2002 and 2022. Only programs that existed throughout the entire study period were kept to ensure comparable data availability across cohorts. The resulting dataset includes 45 undergraduate programs distributed across five campuses, whereas UEM currently offers 64 on-campus undergraduate programs across six campuses. The study includes only undergraduate education; graduate programs are not represented. Throughout the paper, the term *undergraduate program* refers to a specific degree program, such as Medicine or Computer Science.

The dataset contains static attributes available at or before enrollment, including demographic information (e.g., age, gender, race (race categories follow the Brazilian census classification, provided by the Brazilian Institute of Geography and Statistics (IBGE)), and place of birth), admission information (e.g., admission process and quota status), and institutional information (e.g., undergraduate program, campus, and academic center).

After the preprocessing steps described in Section 2.2, the modeling dataset contained 66,820 student records. Table 1 summarizes the characteristics of this population by dropout status. All predictors except age at enrollment are categorical and are therefore reported as category frequencies. For descriptive purposes, age at enrollment is reported in its original continuous form as mean [min, max] and median [Q1, Q3]. Both summaries are provided because its distribution is right-skewed, with most students entering higher education at a typical age but a smaller number of older students extending the distribution to 66 years. Standardized mean differences (SMDs) are reported to quantify differences between the dropout and non-dropout groups. The p-values were obtained using the chi-squared test for categorical variables and the Kruskal-Wallis test for age.

Table 1. Baseline characteristics of the study population, stratified by dropout status.

Variable	Category	Overall (n = 66,820)	Non-Dropout (n = 42,487)	Dropout (n = 24,333)	SMD	p
Age at enrollment	mean [min, max]	19.8 [15.0, 66.0]	19.6 [15.0, 66.0]	20.1 [16.0, 65.0]	−0.112	<0.001
	median [Q1, Q3]	18.0 [17.0, 20.0]	18.0 [17.0, 20.0]	18.0 [17.0, 21.0]	−0.112	<0.001
Study shift	Afternoon	735 (1.1)	282 (0.7)	453 (1.9)	0.221	<0.001
	Afternoon/evening	341 (0.5)	150 (0.4)	191 (0.8)		
	Evening	28,172 (42.2)	16,748 (39.4)	11,424 (46.9)		
	Full−time	26,789 (40.1)	18,360 (43.2)	8429 (34.6)		
	Morning	9810 (14.7)	6384 (15.0)	3426 (14.1)		
	Morning/afternoon/evening	973 (1.5)	563 (1.3)	410 (1.7)		
Marital status	Married	3205 (4.8)	1875 (4.4)	1330 (5.5)	0.112	<0.001
	Not declared	553 (0.8)	218 (0.5)	335 (1.4)		
	Other	144 (0.2)	78 (0.2)	66 (0.3)		
	Separated/divorced	1242 (1.9)	709 (1.7)	533 (2.2)		
	Single	61,662 (92.3)	39,597 (93.2)	22,065 (90.7)		
	Widowed	14 (0.0)	10 (0.0)	4 (0.0)		
Race	Asian	2909 (4.4)	1998 (4.7)	911 (3.7)	0.184	<0.001
	Black	1484 (2.2)	945 (2.2)	539 (2.2)		
	Indigenous	150 (0.2)	70 (0.2)	80 (0.3)		
	Mixed (pardo)	7969 (11.9)	4945 (11.6)	3024 (12.4)		
	Not declared	17,279 (25.9)	9853 (23.2)	7426 (30.5)		
	White	37,029 (55.4)	24,676 (58.1)	12,353 (50.8)		
Admission process	Other	3285 (4.9)	2358 (5.5)	927 (3.8)	0.122	<0.001
	PAS	5683 (8.5)	3503 (8.2)	2180 (9.0)		
	Remaining vacancies	1405 (2.1)	705 (1.7)	700 (2.9)		
	SISU	369 (0.6)	192 (0.5)	177 (0.7)		
	Vestibular	56,078 (83.9)	35,729 (84.1)	20,349 (83.6)		
Gender	Female	35,089 (52.5)	23,537 (55.4)	11,552 (47.5)	0.159	<0.001
	Male	31,731 (47.5)	18,950 (44.6)	12,781 (52.5)		
High school type	Abroad	52 (0.1)	30 (0.1)	22 (0.1)	0.114	<0.001
	Private equivalency	251 (0.4)	112 (0.3)	139 (0.6)		
	Private, throughout	26,153 (39.1)	16,391 (38.6)	9762 (40.1)		
	Public equivalency	678 (1.0)	343 (0.8)	335 (1.4)		
	Public, partly	3363 (5.0)	2134 (5.0)	1229 (5.1)		
	Public, throughout	33,237 (49.7)	21,731 (51.1)	11,506 (47.3)		
	Public, unspecified	3086 (4.6)	1746 (4.1)	1340 (5.5)		
Quota admission	No	60,840 (91.1)	39,100 (92.0)	21,740 (89.3)	0.092	<0.001
	Yes	5980 (8.9)	3387 (8.0)	2593 (10.7)		
Residence region	Campus city	31,304 (46.8)	20,999 (49.4)	10,305 (42.3)	0.165	<0.001
	Other cities in Paraná	24,456 (36.6)	15,177 (35.7)	9279 (38.1)		
	Other states	1660 (2.5)	1049 (2.5)	611 (2.5)		
	State of São Paulo	9400 (14.1)	5262 (12.4)	4138 (17.0)		
Birth region	Campus city	19,001 (28.4)	13,232 (31.1)	5769 (23.7)	0.183	<0.001
	Other cities in Paraná	30,029 (44.9)	18,788 (44.2)	11,241 (46.2)		
	Other states	4505 (6.7)	2780 (6.5)	1725 (7.1)		
	State of São Paulo	13,285 (19.9)	7687 (18.1)	5598 (23.0)		
Disability status	No	65,671 (98.3)	41,772 (98.3)	23,899 (98.2)	0.008	0.351
	Yes	1149 (1.7)	715 (1.7)	434 (1.8)		

The undergraduate program variable has 45 categories. Program sizes range from 362 to 3682 students, with a median of 985, and program-level dropout rates range from 11.0% in Medicine to 68.0% in Mathematics.

Table 1 also shows that several variables are highly concentrated. Most students are recorded as single (92.3%), were admitted outside the quota system (91.1%), and have no registered disability (98.3%). Disability status is the only variable with no statistically significant difference between the dropout and non-dropout groups ($p = 0.351$). Although the remaining variables show statistically significant differences, all standardized mean differences are below 0.25. Overall, when considered individually, the enrollment-time characteristics show limited differences between the two groups.

2.2. Data Preprocessing and Feature Engineering

The raw institutional dataset contained missing values, inconsistent categorical entries, and incomplete academic records, which were addressed through a series of preprocessing steps.

Missing values were treated according to the meaning of each attribute. For categorical variables in which missing information represented a valid institutional state, missing entries were retained as an explicit *not declared* category. This category occurs in two predictors: race and marital status. When missing information could be reliably recovered from other institutional variables, the corresponding value was completed. For example, missing geographic information was inferred only when nationality, state of birth, or other auxiliary attributes provided unambiguous evidence. Admission-related fields were similarly completed using more detailed institutional records. Records with missing or inconsistent values that could not be reliably resolved were excluded.

After preprocessing, the modeling dataset contained no missing values. The effect of retaining *not declared* as an explicit category was evaluated through the sensitivity analysis described in Section 2.9.

Date variables were transformed into derived attributes, including year of birth and age at enrollment. Age at enrollment was subsequently discretized into ordered bands for modeling; the resulting categories are listed in Appendix A. Table 1 reports age in its original continuous form for descriptive purposes. Geographic variables were grouped into broader regional categories.

Program-level information, including degree type, academic center, campus, and maximum program duration, was obtained from official university sources and merged with the student-level data, using the undergraduate program code as the integration key.

The target variable was derived from the latest academic status recorded for each student. A binary variable, *dropout status*, was created to distinguish students who dropped out from those who did not. Students who graduated, remained enrolled, or were dismissed after reaching the maximum time allowed for degree completion were assigned to the *non-dropout* class, whereas students whose enrollment was cancelled or who transferred to another program were assigned to the *dropout* class.

To ensure consistency with an early prediction setting, only variables available at or before enrollment were used as predictors. Variables containing future information, such as the date of the latest academic status or the elapsed time until that status, were excluded from model training. The cohort definition and generation-completeness criterion are described in Section 2.4. These restrictions reduce the risk of information leakage and ensure that the models use only information that would be available at deployment.

The complete set of predictors, including demographic, admission-related, geographic, program-level, and engineered enrollment-time variables, is provided in Appendix A.

2.3. Association Analysis

All predictors except age at enrollment are nominal categorical variables, as is the binary target. Age at enrollment is represented by ordered categories. Because conventional correlation coefficients are not appropriate for nominal variables, associations among predictors and between predictors and the binary dropout target were measured using Cramér's V. This chi-squared-based measure ranges from 0 to 1, where 0 indicates no association and 1 indicates perfect association. We used the bias-corrected version proposed by Bergsma [23], which reduces the upward bias that can occur for variables with many categories. This correction is particularly relevant for the undergraduate program variable, which has 45 categories.

For the association analysis, the age bands were treated as categories so that Cramér's V could be used for all variable pairs without assuming a monotonic relationship between age and the other variables. As a complementary analysis using age in its original continuous form, differences between the dropout and non-dropout groups were assessed using the Kruskal–Wallis test and the standardized mean difference, as reported in Table 1.

2.4. Cohort Definition and Generation Completeness

To account for the temporal structure of undergraduate programs, students were grouped into *generations*. A generation is defined as the cohort of students admitted to the same undergraduate program in the same year. Formally, for a program p and an admission year y, the corresponding generation is

$$G_{p,y} = \{\, s \in S \;:\; \text{program}(s) = p \;\wedge\; \text{year}(s) = y \,\}, \tag{1}$$

where S denotes the set of all student records in the dataset. Because the dataset spans admissions between 2002 and 2022 across 45 programs, this grouping partitions the records into program-specific annual cohorts that share the same curriculum and institutional conditions.

Generations differ in whether sufficient time has elapsed for their academic outcomes to be observed. Each program p has a maximum allowed duration for degree completion, denoted $D_{\max}(p)$ and obtained from official university records (maximum program durations were merged into the student-level data using the undergraduate program code as the integration key, as described in Section 2.2). Letting $y_{\text{obs}} = 2022$ denote the last year covered by the dataset, a generation is classified as *complete* when

$$y + D_{\max}(p) \;\leq\; y_{\text{obs}} \tag{2}$$

and as *incomplete* otherwise. Under this criterion, students in a complete generation have been observed for at least the maximum duration allowed for their program, providing sufficient time for a terminal academic status to occur.

The Computer Science program illustrates this criterion. Its minimum duration is four years, and its maximum allowed duration is eight years. Students admitted in 2014 reach the end of the maximum allowed period in 2022, making $G_{\text{CS},2014}$ a complete generation. In contrast, students admitted in 2015 reach this limit only in 2023. Thus, $G_{\text{CS},2015}$ and subsequent Computer Science generations are incomplete within the observation window, even though some students in these generations may already have graduated or left the program.

This distinction is important because incomplete generations are subject to right-censoring. Students recorded as still *enrolled* have not yet reached a terminal status and may drop out after the end of the observation window. Assigning these students to the *non-dropout* class is a provisional labeling decision rather than an observed final outcome,

introducing label noise primarily in incomplete generations. In the modeling dataset, 8645 of the 8787 students with an *enrolled* status (98.4%) belong to incomplete generations, whereas only 142 belong to complete generations.

Applying the completeness criterion to the modeling dataset yields 42,249 records in complete generations and 24,571 records in incomplete generations. Both subsets are retained for the experimental analyses. The experimental design described in Section 2.5 compares the use of all available records with the use of complete generations only, allowing the effect of potentially censored outcomes on predictive performance to be assessed.

2.5. Experimental Design and Training Scenarios

To evaluate model behavior under different institutional and temporal conditions, five training scenarios were defined:

1. Training using the complete dataset.
2. Training using only complete generations.
3. Training on older complete generations and testing on more recent complete generations.
4. Training and evaluation within individual academic centers.
5. Training and evaluation within individual undergraduate programs.

In Scenario 1, all 66,820 student records were used regardless of generation completeness. Scenario 2 was restricted to the 42,249 records belonging to complete generations.

Scenario 3 used a temporal split in which models were trained on complete generations from the ten oldest admission years and tested on complete generations from the six most recent admission years. This configuration simulates a realistic deployment setting in which historical data are used to predict outcomes for future student cohorts [16,21]. The resulting partition contained 31,946 training records and 10,303 test records.

To assess whether the results of Scenario 3 depended on the specific temporal cut, an additional rolling-origin analysis was conducted across the admission years represented by complete generations. The cut between training and test cohorts was progressively advanced from the sixth to the twelfth admission year, yielding seven temporal partitions. Earlier cuts were excluded because the training set would be smaller than the corresponding test set, whereas later cuts were excluded because fewer than 3000 students would remain in the test set. For each cut, the models were refitted using the hyperparameters selected for Scenario 3 and evaluated on the subsequent cohorts.

Each rolling-origin partition was evaluated at two decision thresholds. The first was the fixed threshold of 0.5 used for the main performance results. The second was a prevalence-matched threshold that classified as *dropout* the highest-risk proportion of test students corresponding to the dropout prevalence observed in that partition's training cohorts. Thus, threshold selection used the dropout prevalence from the training data and the predicted probabilities for the test data, without using test-set outcome labels. Recall and F1-score for the *dropout* class were summarized across the seven temporal cuts using their mean, standard deviation, and range.

To investigate context-specific predictive behavior, Scenario 4 trained and evaluated models separately within two academic centers selected according to their class distributions. The Center of Technology (CTC) was selected as representative of the class-distribution pattern observed in most academic centers, where *non-dropout* students outnumber *dropout* students. It is also among the largest centers in the dataset. The Center of Exact Sciences (CES) was selected as a contrasting case because it is the only academic center for which *dropout* students outnumber *non-dropout* students. Models were not trained separately for the remaining academic centers, as their class distributions were similar to that observed for the CTC.

Scenario 5 provided a more fine-grained analysis using three undergraduate programs selected to represent different class distributions. Computer Science and Languages have relatively balanced class distributions, whereas Information Technology has a slightly higher proportion of *dropout* students.

Because the program-level datasets are relatively small, the stability of the results was additionally assessed using 50 repeated stratified 80/20 train–test splits for each program. The three tree-based models were refitted in each split using the hyperparameters previously selected for that program, so that the analysis measured variability at a fixed model configuration rather than variability introduced by repeated hyperparameter selection. The logistic regression baseline described in Section 2.6 re-selected its regularization strength within each split. Mean *dropout* recall and F1-score were computed across the 50 repetitions.

Because training sets from repeated random splits overlap, conventional standard errors would underestimate uncertainty; therefore, the 95% confidence intervals were computed using the corrected variance estimator of Nadeau and Bengio [24], which accounts for dependence induced by overlapping training sets.

Except for Scenario 3, each dataset was divided into training and test sets using a stratified 80/20 split to preserve the class distribution. Scenario 3 instead used the temporal partition described above. Hyperparameters were selected by grid search with 5-fold cross-validation using only the training data [25]. The best-performing configurations were then evaluated on the held-out test set.

All scenarios were derived from the same modeling dataset, so the center-level and program-level analyses correspond to subsets of this dataset rather than to separate data sources. Each student belongs to a single academic center and undergraduate program, and the predictors were constructed consistently across all scenarios.

Table 2 summarizes the data partitions used in each scenario. The training and test sets are disjoint at the student level. For the stratified random splits, dropout prevalence is approximately preserved between the training and test sets. Scenario 3 differs by design because its partition follows admission year; dropout prevalence is 0.33 in the training cohorts and 0.43 in the test cohorts.

Table 2. Data partitions used in the five training scenarios. Prevalence denotes the proportion of students in the *dropout* class.

				Prevalence	
Scenarios	**Students**	**Training**	**Test**	**Training**	**Test**
Scenario 1: Complete dataset	66,820	53,456	13,364	0.36	0.36
Scenario 2: Complete generations	42,249	33,799	8450	0.36	0.36
Scenario 3: Temporal split	42,249	31,946	10,303	0.33	0.43
Scenario 4: CTC	15,813	12,650	3163	0.37	0.37
Scenario 4: CES	6229	4983	1246	0.60	0.60
Scenario 5: Computer Science	985	788	197	0.43	0.43
Scenario 5: Languages	3318	2654	664	0.44	0.44
Scenario 5: Information Technology	949	759	190	0.52	0.53

2.6. Machine Learning Models

Three supervised ML classifiers were used as the primary models throughout the experimental scenarios: Decision Tree (DT) [26], Random Forest (RF) [27], and XGBoost [28]. The task was formulated as the binary classification problem of predicting whether a student belongs to the *dropout* or *non-dropout* class.

DT, RF, and XGBoost were selected because they are well suited to structured tabular data, can capture nonlinear relationships, and are widely used in academic dropout prediction [13,16,29,30]. DT provides interpretable decision rules, RF reduces variance by combining multiple trees [16,29], and XGBoost uses regularized gradient boosting and has shown competitive performance on structured educational data [13,30].

Categorical predictors were label-encoded for DT, RF, and XGBoost. This representation was used because tree-based models partition the feature space through split rules, making them less directly affected by the ordinal interpretation of integer codes than linear or fully connected models. The implications of this encoding for SMOTE-NC are addressed separately in Section 2.7.

Two additional models were used as reference baselines: regularized Logistic Regression (LR) and a feed-forward Neural Network (NN). LR was evaluated on the complete dataset and temporal split, and was also included in the repeated program-level analysis. The NN was evaluated on the complete dataset and temporal split. These baselines were used to assess whether the performance patterns observed for the tree-based models were specific to the modeling family. The NN contained two hidden layers with 64 and 32 units and was trained with early stopping. Both baselines used one-hot encoded nominal predictors, while the age band was retained as an ordinal variable and standardized. This representation avoids imposing an arbitrary numeric ordering on nominal categories, which would be inappropriate for linear and fully connected models.

Hyperparameters were selected using exhaustive grid search with 5-fold stratified cross-validation on the training data. All parameter combinations listed in Table 3 were evaluated across the five folds, with no early-stopping criterion for the grid search. The search comprised 90 configurations for DT, 216 each for RF and XGBoost, four for LR, and six for the NN. The configuration with the highest mean cross-validated accuracy was selected for each model. The selected models were subsequently evaluated on the held-out test set using the metrics described in Section 2.10.

Table 3. Hyperparameter search spaces for the evaluated models.

Model	Hyperparameter	Values	Combinations
Decision Tree	criterion max_depth min_samples_split min_samples_leaf	gini, entropy 3, 5, 10, 20, None 2, 10, 20 1, 5, 10	90
Random Forest	n_estimators max_depth min_samples_split min_samples_leaf max_features	100, 200, 300 10, 20, 30, None 2, 5, 10 1, 2, 4 sqrt, log2	216
XGBoost	n_estimators max_depth learning_rate subsample colsample_bytree min_child_weight	100, 200, 300 3, 5, 7 0.05, 0.1, 0.2 0.8, 1.0 0.8, 1.0 1, 5	216
Logistic Regression	C	0.01, 0.1, 1, 10	4
Neural Network	α Initial learning rate	10^{-4}, 10^{-3}, 10^{-2} 10^{-3}, 10^{-2}	6

All models were implemented in Python 3.14.7 using the scikit-learn (version 1.8.0) and xgboost (version 3.2.0) libraries.

2.7. Class Imbalance Handling

To assess the effect of class imbalance, two additional training configurations were derived from Scenario 1 by applying oversampling and undersampling to its training partition. Together with the eight configurations defined by the five training scenarios in Section 2.5, these resampled configurations yield the ten training configurations evaluated for the three primary models. Resampling was performed using the imbalanced-learn library (version 0.14.1) [31]:

- Oversampling: SMOTE-NC [32] was used to generate synthetic instances of the minority class until both classes were equally represented.
- Undersampling: Majority-class instances were randomly removed until both classes were equally represented.

Both resampling configurations used the same train/test partition as Scenario 1. Resampling modified only the training partition. The original 53,456 training records increased to 67,980 after oversampling and decreased to 38,932 after undersampling, while the test partition remained unchanged with 13,364 records and the original class distribution.

SMOTE-NC is designed for datasets containing both nominal and continuous predictors [32]. It one-hot encodes the nominal predictors internally before the nearest-neighbor search, so the integer codes assigned to categories are not treated as numeric magnitudes, and it draws the nominal values of each synthetic instance from the categories observed among its neighbors rather than interpolating them. As a result, synthetic instances cannot contain invalid categorical values, and the neighbor structure does not depend on the integer codes.

The age-band predictor was the only variable not designated as categorical in SMOTE-NC. Its ordered integer representation was treated numerically during neighbor construction and interpolation, reflecting the ordinal structure of the age bands.

Resampling was performed within each cross-validation fold rather than on the training partition before cross-validation. The resampler and classifier were combined into a single pipeline passed to the grid search. Within each fold, the resampler was fitted only on the training portion, while the validation portion retained its original class distribution. This procedure prevents observations from the validation fold from contributing to synthetic or resampled training instances and avoids information leakage during model selection. After hyperparameter selection, the final model was fitted using the resampled full training partition and evaluated on the unchanged test partition.

To assess whether the oversampling results were sensitive to the arbitrary integer codes assigned to nominal predictors, the complete oversampling experiment was additionally repeated under three independent random relabelings of all nominal categories. Pairwise distances in the internal SMOTE-NC representation were compared across relabelings, and the three classifiers were refitted using the same hyperparameters as in the original oversampling configuration. Changes in *dropout* recall and PR-AUC were then compared with the gain obtained from resampling itself.

Because the unresampled, oversampled, and undersampled models were evaluated on the same test partition, pairwise differences between their predictions were assessed using McNemar's test. The test was applied both to the overall correctness and separately to the subset of students who actually dropped out, with the latter providing a paired comparison of *dropout* recall.

The effect of oversampling and the sensitivity of the results to the encoding strategy are examined empirically in Section 3.4.

2.8. Preprocessing Pipeline and Leakage Controls

The preprocessing pipeline outlined in Figure 1 was executed in a fixed order. Raw institutional records were first cleaned and standardized, with missing information either retained as an explicit *not declared* category, recovered from auxiliary attributes when unambiguous, or excluded when it could not be reliably resolved. Enrollment-time predictors were then constructed, generations were assigned according to the completeness criterion described in Section 2.4, and program-level attributes were merged. The resulting data were partitioned into training and test sets before any model-dependent transformations were fitted.

Model-specific encoding, resampling when applicable, and hyperparameter selection were performed using only the training data. During cross-validation, all fitted preprocessing and resampling operations were restricted to the training portion of each fold, leaving the corresponding validation portion unchanged. The held-out test set was not used for model selection or preprocessing estimation and was evaluated only after the final model had been fitted. Table 4 summarizes the main leakage controls.

In Scenarios 1, 2, 4, and 5, the stratified 80/20 split mixes admission years across the training and test partitions. This does not constitute record-level data leakage because the partitions are disjoint at the student level, but it also does not enforce the temporal ordering expected in a deployment setting. Scenario 3 addresses this limitation by training exclusively on older complete generations and testing on more recent complete generations. The effects of this temporal separation are examined in Sections 3.3.3 and 3.5.

Table 4. Summary of preprocessing steps and data leakage controls.

Step	Data Used	Leakage Control
Missing-value handling	Individual record	Rule-based treatment using the student's own attributes; no statistics estimated from other records.
Predictor construction	Individual record	Deterministic transformations using information available at or before enrollment.
Predictor selection	Not fitted	Post-enrollment variables and variables used only for cohort construction excluded from model inputs.
Generation completeness	Not fitted	Used only to construct experimental subsets; never supplied to the models as a predictor.
Categorical encoding	Training data	Encoding fitted without using information from the held-out test set; label encoding used for tree-based models and one-hot encoding for the reference baselines.
Class imbalance handling	Training folds	SMOTE-NC or random undersampling applied only within training data; validation and test distributions left unchanged.
Hyperparameter selection	Training data	Five-fold cross-validation restricted to the training partition; held-out test set evaluated only after model selection.

2.9. Sensitivity Analysis of Undeclared Values

As described in Section 2.2, missing information representing a valid institutional state was retained as an explicit *not declared* category for race and marital status rather than being imputed. To assess whether this treatment affected the predictive results, a sensitivity analysis was conducted using the complete-dataset and temporal-split configurations.

For each configuration, the three primary models were refitted under three alternative treatments of *not declared* values:

1. Mode imputation: Each *not declared* value was replaced with the most frequent observed category of the corresponding predictor in the training partition.
2. Removal from training: Records containing *not declared* values for either race or marital status were removed from the training partition, while the test partition was left unchanged.
3. Removal from both partitions: Records containing *not declared* values for either race or marital status were removed from both the training and test partitions.

The second treatment preserves the original test population, allowing direct comparison with the results obtained when *not declared* is retained as a category. The third treatment also removes undeclared values from the test data, but changes the evaluated population; thus, it is used to compare performance trends rather than directly-matched metric values.

The models were refitted using the hyperparameters previously selected for the corresponding configuration, ensuring that the sensitivity analysis isolated the effect of the treatment of undeclared values rather than repeating model selection. Appendix C compares the resulting performance with the original treatment in which *not declared* is retained as an explicit category.

2.10. Model Evaluation

Model performance was evaluated using accuracy, precision, recall, and F1-score. Because the task was formulated as a binary classification problem, these metrics were computed from the confusion matrix. Taking the *dropout* class as the positive class, true positives (TP) are students who dropped out and were correctly classified as such, false negatives (FN) are students who dropped out but were classified as *non-dropout*, true negatives (TN) are correctly classified *non-dropout* students, and false positives (FP) are *non-dropout* students incorrectly classified as *dropout*. The metrics are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{5}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

Accuracy is a model-level metric, whereas precision, recall, and F1-score depend on which class is treated as positive and as such are reported separately for each class. For the *dropout* class, these metrics follow Equations (4)–(6); for the *non-dropout* class, the roles of the two classes are reversed. Accordingly, the result tables report the per-class precision, recall, and F1-score together with a single accuracy value for each model.

Given the objective of identifying students at risk of dropout, recall for the *dropout* class was considered the primary metric for evaluating the selected models. It measures the proportion of actual dropout cases that were correctly identified, whereas precision measures the proportion of students predicted as *dropout* who actually dropped out. The F1-score provides a combined measure of precision and recall. For context, the held-out test partition of the complete dataset contains 4867 students in the *dropout* class (Table 2); therefore, one percentage point of *dropout* recall corresponds to approximately 49 students correctly identified or missed by the model.

Because the *dropout* class is less prevalent, accuracy alone can be misleading: in the complete dataset, a classifier that predicted every student as *non-dropout* would reach 63.58%

accuracy with zero recall for dropout. False negatives are the most consequential error in this application, because students who are not identified cannot be offered early support.

For the main results of the five training scenarios, accuracy, precision, recall, and F1-score were computed from class predictions obtained using the default decision threshold of 0.5. In the rolling-origin analysis of Scenario 3, recall and F1-score were additionally evaluated using the prevalence-matched threshold defined in Section 2.5. This complementary evaluation was used to examine the effect of the operating threshold on dropout identification under temporal distribution shift.

Because performance at a single decision threshold does not fully characterize the predicted probabilities, PR-AUC and ROC-AUC were also evaluated. PR-AUC summarizes the precision-recall trade-off across thresholds and is particularly informative under class imbalance. Its baseline corresponds to the prevalence of the positive class; therefore, dropout prevalence is reported alongside PR-AUC together with the ratio between PR-AUC and prevalence, referred to as *lift*. ROC-AUC is included to facilitate comparison with the broader literature.

The Brier score was used to assess the probabilistic prediction error:

$$\text{Brier} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2 \tag{7}$$

where $\hat{p}_i$ is the predicted probability of dropout for student i and $y_i \in \{0,1\}$ is the observed outcome. Lower Brier scores indicate smaller discrepancies between predicted probabilities and observed outcomes. Calibration was further examined using reliability diagrams, which compare the mean predicted probability within each risk bin with the observed dropout frequency in that bin. These probability-based measures complement the classification metrics by evaluating the models' ability to rank students by risk and to provide predicted probabilities that reflect observed dropout frequencies.

2.11. Fairness and Institutional Bias Analysis

Predictive performance measured over the whole population does not reveal whether a model behaves equitably across different student groups therefore, we evaluated how the model's predictions and errors are distributed across student groups. The analysis was carried out for the best-performing model, XGBoost, on the complete-dataset scenario (Scenario 1). It used the same held-out test partition and the same predictions as the main results, taken at the default probability threshold of 0.5, making the group-level counts consistent with the confusion matrix reported for that scenario. No additional model was trained for this analysis.

We examined five attributes available at enrollment: gender, self-declared race, age at enrollment, quota admission status, and high school type. These were selected because they describe demographic and socioeconomic groups that an early warning system would directly affect. Race and gender used the categories listed in Table 1, high school type was grouped into public or private, and age at enrollment was grouped into three bands: up to 18 years, 19 to 24 years, and 25 years or older.

For each group, we report the observed dropout rate (base rate), the proportion of students flagged as *dropout* by the model (selection rate), the recall for the *dropout* class (true positive rate), the false positive rate, the mean predicted dropout probability, and the ROC-AUC. Recall measures how often students who actually dropped out are identified, and equality of recall across groups corresponds to the equal opportunity criterion; reporting the false positive rate alongside it corresponds to the equalized odds criterion. Comparing the mean predicted probability with the observed dropout rate provides an assessment of the overall calibration within each group. The main summary

for each attribute is the difference between the largest and smallest group values of recall and false positive rate.

2.12. Model Interpretation

To examine which predictors contributed most strongly to model predictions, two complementary feature attribution approaches were used. First, the feature importance scores provided by the tree-based models were analyzed. For Decision Tree and Random Forest, importance was based on impurity reduction, whereas for XGBoost it was based on gain. Because these measures are defined differently across model families, their magnitudes were not compared directly between models; instead, the analysis focused on the relative ranking of predictors within each model.

Feature importance was examined for selected models and experimental configurations. For the complete-dataset and temporal-split configurations, Random Forest and XGBoost were analyzed to assess whether the relative importance of predictors changed under temporal separation. At the program level, feature importance was reported for the model achieving the highest *dropout* recall in each program.

Because impurity- and gain-based importance measures can be affected by predictor cardinality and dependencies among predictors, SHapley Additive exPlanations (SHAP) [33] was additionally used to examine feature contributions on held-out data as a complement to the internal importance measures.

SHAP values were obtained with the TreeExplainer of the shap library (version 0.52.0) [33], which computes exact Shapley values for tree ensembles. We used path-dependent tree traversal, so no separate background or reference dataset was required, and we computed all values for the *dropout* (positive) class. Global importance was summarized as the mean absolute SHAP value of each predictor and expressed as a percentage of the total across predictors, allowing models with different raw scales to be compared. The figures in Section 3.7.1 report these values for Random Forest and XGBoost under the complete-dataset and temporal-split scenarios.

3. Results and Discussion

This section presents and discusses the main results of the study, including descriptive and association analyses, model performance across the experimental configurations, probability-based evaluation and calibration, model interpretation, and sensitivity analyses.

3.1. Descriptive Analysis of Academic Dropout

The modeling dataset was composed of 66,820 students, 24,333 (36.42%) classified as *dropout* and 42,487 (63.58%) as *non-dropout*. This is a moderate class imbalance, with *non-dropout* students accounting for nearly two thirds of the records. The dropout proportion is 0.36 in both the complete-dataset and complete-generation configurations and rises to 0.43 in the test partition of the temporal split (Table 2).

Table 5 shows the distribution of the original academic status categories and their mapping to the binary target. Most students either graduated (48.71%) or had their enrollment cancelled (35.32%). Cancelled enrollments account for 23,603 of the 24,333 *dropout* cases (97.0%), while transfers account for the remaining 3.0%. Transfers represent program-level rather than institutional dropout, consistent with the program-centered definition adopted in this study.

Figure 2 shows that dropout events are concentrated in the early stages of enrollment, the largest number occurs within the first six months, with progressively fewer cases thereafter. Most dropout occurs before academic performance records would accumulate,

which again supports our aim of predicting dropout risk using only information available at enrollment.

Table 5. Mapping between the original academic status categories and the binary target variable.

Academic Status	N	%	Class
Graduated	32,549	48.71	Non-dropout
Enrolled	8787	13.15	Non-dropout
Dismissed (time limit)	1151	1.72	Non-dropout
Cancelled	23,603	35.32	Dropout
Transferred	730	1.10	Dropout
Total	66,820	100.00	

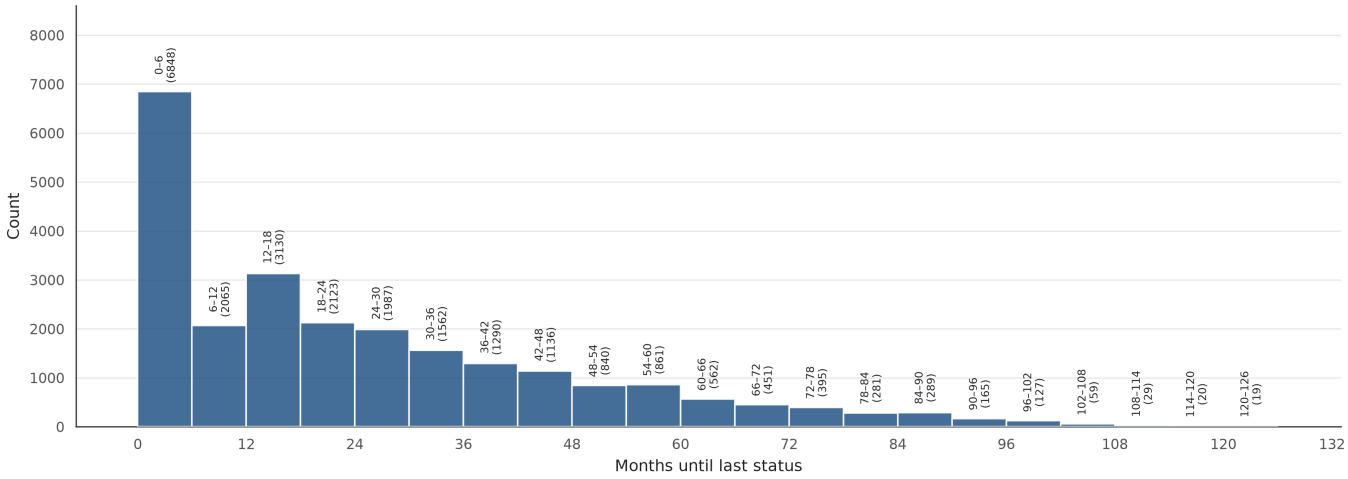


Figure 2. Distribution of dropout cases according to time enrolled (in months) until dropout.

3.2. Association Analysis

Figure 3 presents the association matrix for the twelve predictors and the binary target, computed using all 66,820 records in the modeling dataset as described in Section 2.3.

Associations between individual predictors and dropout status are generally weak: undergraduate program has the strongest association with the target ($V = 0.27$); all other predictors have values of $V \leq 0.11$, with study shift ($V = 0.11$), race ($V = 0.09$), and birth region ($V = 0.09$) showing the largest associations among them. Disability status shows no association with dropout ($V = 0.00$), consistent with the non-significant difference reported in Table 1.

The strongest associations occur between predictors: undergraduate program is strongly associated with study shift ($V = 0.68$) and moderately associated with gender ($V = 0.42$). Residence region is strongly associated with birth region ($V = 0.58$), while high school type is associated with quota status ($V = 0.29$). These relationships are consistent with characteristics of the dataset and the institutional context: programs are offered in specific study shifts, students often study in the region where they were born, and quota eligibility is partly determined by prior schooling.

Because related predictors may share predictive information, these associations should be considered when interpreting feature importance. Section 3.7.1 further examines predictor importance using both tree-based importance measures and SHAP values.

Figure 3. Association matrix (bias-corrected Cramér's V) among the twelve enrollment-time predictors and the binary dropout target.

3.3. Model Performance Across Training Scenarios

This section reports the performance of the three models across the training scenarios described in Section 2.5. The main text covers the three institution-wide configurations: the complete dataset (Scenario 1), the complete generations (Scenario 2), and the temporal split (Scenario 3). All results were obtained on the held-out test partitions summarized in Table 2, using the default decision threshold of 0.5.

The detailed results for academic centers and individual programs (Scenarios 4 and 5), including the repeated-split analysis, are reported in Appendices B.1 and B.2. Finally, Appendix B.3 compares the results across all training configurations.

3.3.1. Scenario 1: Complete Dataset

Table 6 presents the performance of the three models on the complete dataset.

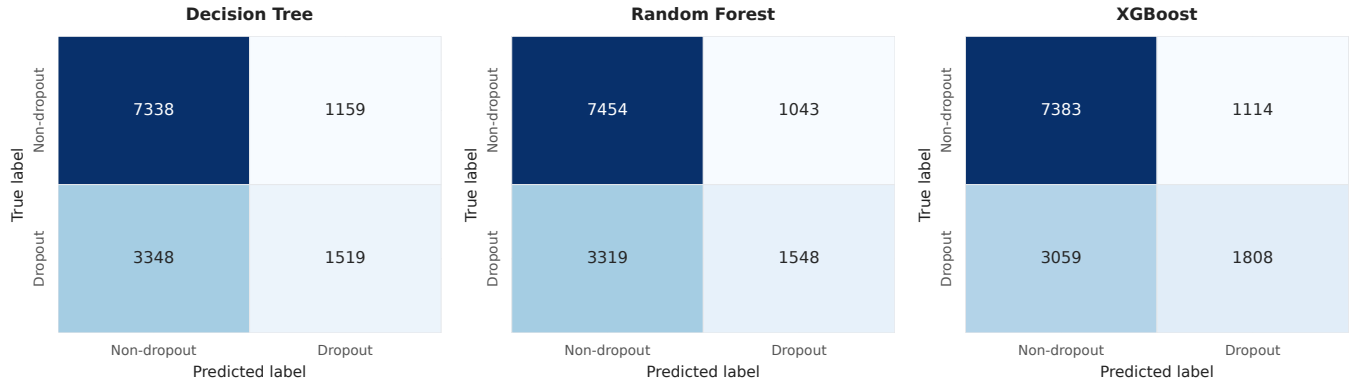
Table 6. Performance of the models on the complete dataset (Scenario 1).

Model	Class	Precision	Recall	F1-Score	Accuracy
Decision Tree	Non-dropout	0.69	0.86	0.77	0.66
	Dropout	0.57	0.31	0.40	
Random Forest	Non-dropout	0.69	0.88	0.77	0.67
	Dropout	0.60	0.32	0.42	
XGBoost	Non-dropout	0.71	0.87	0.78	0.69
	Dropout	0.62	0.37	0.46	

XGBoost achieved the best performance among the three models, with the highest *dropout* recall (0.37), *dropout* F1-score (0.46), and overall accuracy (0.69). Nevertheless, *dropout* recall remained low for all models (0.31–0.37), indicating that most dropout cases were not identified at the default decision threshold.

Figure 4 shows the corresponding confusion matrices and makes this error pattern explicit. False negatives are the most frequent error for all three models: Decision Tree

misses 3348 of the 4867 dropout cases, Random Forest 3319, and XGBoost 3059. Compared with Random Forest, XGBoost identifies 260 additional dropout students at the cost of 71 additional false positives. At the default threshold, XGBoost would flag 2922 of the 13,364 students in the test partition (21.9%), of whom 1808 (61.9%) actually dropped out. These results illustrate the practical trade-off between identifying more at-risk students and increasing the number of false alarms.



(**a**) Decision Tree (**b**) Random Forest (**c**) XGBoost

Figure 4. Confusion matrices for the three models trained on the complete dataset (Scenario 1), evaluated on the held-out test partition (n = 13,364).

3.3.2. Scenario 2: Complete Generations

Table 7 presents model performance when training and evaluation are restricted to complete generations.

Restricting the analysis to complete generations produced modest improvements over Scenario 1 for all three models. *Dropout* recall increased from 0.31 to 0.34 for Decision Tree, from 0.32 to 0.35 for Random Forest, and from 0.37 to 0.41 for XGBoost. The corresponding F1-scores increased from 0.40 to 0.43, from 0.42 to 0.45, and from 0.46 to 0.49, respectively. XGBoost remained the best-performing model, reaching an overall accuracy of 0.70 and a *dropout* recall of 0.41.

These modest improvements are consistent with the reduced exposure to provisional labels caused by right-censoring in incomplete generations. However, the overall performance pattern remains similar to Scenario 1 in that restricting the dataset to complete generations does not eliminate the difficulty of identifying dropout students from enrollment-time predictors alone.

Table 7. Performance of the models using only complete generations (Scenario 2).

Model	Class	Precision	Recall	F1-Score	Accuracy
Decision Tree	Non-dropout	0.70	0.86	0.77	0.68
	Dropout	0.58	0.34	0.43	
Random Forest	Non-dropout	0.71	0.88	0.78	0.69
	Dropout	0.62	0.35	0.45	
XGBoost	Non-dropout	0.72	0.86	0.78	0.70
	Dropout	0.62	0.41	0.49	

3.3.3. Scenario 3: Training on Older Generations and Testing on More Recent Ones

Table 8 presents model performance when training on older complete generations and testing on more recent complete generations. The temporal partition also introduces a

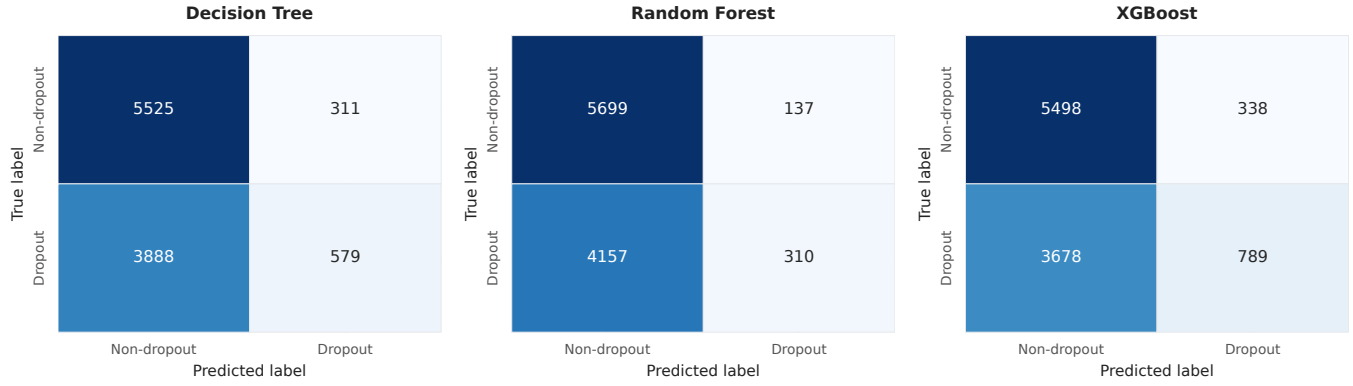
change in class distribution: the proportion of *dropout* students increases from 33.4% in the training cohorts to 43.4% in the test cohorts.

Table 8. Performance of the models when training on older complete generations and testing on more recent complete generations (Scenario 3).

Model	Class	Precision	Recall	F1-Score	Accuracy
Decision Tree	Non-dropout Dropout	0.59 0.65	0.95 0.13	0.72 0.22	0.59
Random Forest	Non-dropout Dropout	0.58 0.69	0.98 0.07	0.73 0.13	0.58
XGBoost	Non-dropout Dropout	0.60 0.70	0.94 0.18	0.73 0.28	0.61

All three models show substantially lower *dropout* recall than in Scenarios 1 and 2. XGBoost performs best for the *dropout* class, with a recall of 0.18 and an F1-score of 0.28, followed by Decision Tree (0.13 and 0.22) and Random Forest (0.07 and 0.13). In contrast, *dropout* precision remains relatively high (0.65–0.70), indicating that the models identify few dropout students but are often correct when they do so.

Figure 5 shows the corresponding confusion matrices. Random Forest classifies only 447 of the 10,303 test students (4.3%) as *dropout* and correctly identifies 310 of the 4467 actual dropout cases. Decision Tree flags 890 students (8.6%) and identifies 579 dropout cases, while XGBoost flags 1127 students (10.9%) and identifies 789. By comparison, the proportions of students classified as *dropout* in Scenario 1 were 20.0%, 19.4%, and 21.9% for Decision Tree, Random Forest, and XGBoost, respectively. Thus, all three models classify substantially fewer students as *dropout* in the newer cohorts, even though the observed dropout prevalence is higher.



(**a**) Decision Tree (**b**) Random Forest (**c**) XGBoost

Figure 5. Confusion matrices for the three models trained on older generations and tested on more recent ones (Scenario 3), evaluated on the held-out recent cohorts ($n = 10{,}303$).

To assess whether this degradation depends on the particular temporal partition, the models were additionally evaluated across seven rolling-origin cuts, as described in Section 2.5. Table 9 summarizes *dropout* recall and F1-score across these cuts at both the default threshold of 0.5 and the prevalence-matched threshold.

At the fixed threshold of 0.5, low *dropout* recall persists across all seven temporal cuts. Mean recall is 0.13 ± 0.05 for Decision Tree, 0.08 ± 0.04 for Random Forest, and 0.17 ± 0.08 for XGBoost, and no model exceeds a recall of 0.29 at any cut. Therefore, the main Scenario 3

split is not an isolated unfavorable partition; its recall values are consistent with the performance observed across the rolling-origin analysis.

Table 9. Dropout recall and F1-score across seven rolling-origin temporal cuts, evaluated at the fixed 0.5 threshold and at the prevalence-matched threshold. Values are means over cuts, with standard deviation and range.

Threshold	Model	Recall	F1-Score
Fixed at 0.5	Decision Tree	0.13 ± 0.05 [0.05, 0.19]	0.21 ± 0.07 [0.10, 0.29]
	Random Forest	0.08 ± 0.04 [0.03, 0.15]	0.13 ± 0.07 [0.06, 0.25]
	XGBoost	0.17 ± 0.08 [0.08, 0.29]	0.26 ± 0.09 [0.14, 0.40]
Prevalence-matched	Decision Tree	0.43 ± 0.03 [0.38, 0.48]	0.48 ± 0.02 [0.46, 0.51]
	Random Forest	0.43 ± 0.02 [0.39, 0.45]	0.49 ± 0.01 [0.48, 0.50]
	XGBoost	0.45 ± 0.02 [0.41, 0.47]	0.52 ± 0.01 [0.51, 0.53]

Performance changes substantially when the operating threshold is adjusted to the dropout prevalence observed in the corresponding training cohorts. Mean *dropout* recall increases to 0.43 for Decision Tree and Random Forest and to 0.45 for XGBoost, with standard deviations of only 0.02–0.03. Mean F1-score similarly increases to 0.48–0.52 with lower variability. The very low recall at the fixed threshold is therefore largely an operating-point effect under temporal distribution shift, a behavior examined further in Section 3.5.

3.4. Impact of Class Imbalance Handling

Table 10 compares model performance after applying oversampling and undersampling to the training partition of Scenario 1.

Table 10. Performance of the models after oversampling and undersampling of the training data from Scenario 1.

Strategy	Model	Class	Precision	Recall	F1-Score	Accuracy
Oversampling	Decision Tree	Non-dropout	0.73	0.64	0.68	0.62
		Dropout	0.48	0.59	0.53	
	Random Forest	Non-dropout	0.74	0.67	0.70	0.64
		Dropout	0.50	0.59	0.54	
	XGBoost	Non-dropout	0.75	0.66	0.70	0.64
		Dropout	0.51	0.61	0.55	
Undersampling	Decision Tree	Non-dropout	0.74	0.63	0.68	0.62
		Dropout	0.48	0.60	0.54	
	Random Forest	Non-dropout	0.75	0.63	0.68	0.63
		Dropout	0.49	0.63	0.55	
	XGBoost	Non-dropout	0.76	0.64	0.70	0.64
		Dropout	0.51	0.64	0.57	

Both resampling strategies substantially increase recall for the *dropout* class relative to the models without resampling in Scenario 1. Oversampling increases *dropout* recall from 0.31–0.37 to 0.59–0.61, while undersampling increases it to 0.60–0.64. This gain is accompanied by lower recall for the *non-dropout* class, illustrating the trade-off between identifying more students who drop out and generating additional false positive classifications.

The two resampling strategies produce similar results. Undersampling yields slightly higher *dropout* recall for all three models, with differences of 0.01 for Decision Tree, 0.04 for Random Forest, and 0.03 for XGBoost; the corresponding F1-score differences do not exceed 0.02. McNemar's test, conducted as described in Section 2.7, provides a more

nuanced comparison. For overall correctness, the differences between oversampling and undersampling are negligible for Decision Tree ($p = 0.88$) and XGBoost ($p = 1.00$) while being small but statistically detectable for Random Forest ($p = 0.04$). When the comparison is restricted to actual *dropout* students, the differences are statistically detectable for all three models ($p = 0.038$, $p = 1.1 \times 10^{-13}$, and $p = 4.4 \times 10^{-9}$ for Decision Tree, Random Forest, and XGBoost, respectively).

The magnitude of these differences is nevertheless small. With 4867 *dropout* students in the common test partition, recall differences of 0.01–0.04 can be statistically detectable without representing a large practical difference between the two strategies; by contrast, comparisons between either rebalanced configuration and the original models without resampling show substantially larger differences ($p < 10^{-17}$ for overall correctness and $p < 10^{-240}$ for *dropout* recall for all three classifiers). Thus, the evidence supports the effect of rebalancing itself more strongly than it supports a substantive advantage of undersampling over oversampling. Whether the recall gains correspond to improved discrimination or primarily to a change in operating point is examined in Section 3.5.

The encoding sensitivity analysis provides additional support for the use of SMOTE-NC. Randomly relabeling the nominal categories leaves pairwise distances in the resampler's internal representation unchanged to floating-point precision, and no synthetic record contains a category that is absent from the observed training data. Across three independent relabelings, *dropout* recall varies by at most 0.044 for Decision Tree, 0.017 for Random Forest, and 0.010 for XGBoost, while PR-AUC varies by at most 0.012, 0.003, and 0.002, respectively. These variations are substantially smaller than the approximately 0.24–0.28 increase in *dropout* recall obtained by oversampling relative to the configuration without resampling, and the ordering of the three models remains unchanged. Therefore, the oversampling gain cannot be attributed primarily to the arbitrary integer codes assigned to nominal categories.

Decision Tree shows the greatest sensitivity to category relabeling. This is consistent with its use of threshold-based splits on integer-encoded nominal predictors, for which relabeling changes the category partitions that can be represented by an individual split. The ensemble models show considerably smaller variation, indicating greater robustness to this residual effect of label encoding.

Among the evaluated configurations, XGBoost with undersampling obtains the highest *dropout* recall (0.64) and F1-score (0.57). However, the small differences between the two resampling strategies and the trade-off with *non-dropout* performance do not support treating this configuration as universally preferable. The threshold-free analysis in Section 3.5 further examines whether rebalancing improves the models' ability to discriminate between students at different levels of dropout risk.

3.5. Threshold-Free Performance and Calibration

The results reported in the preceding sections use the default decision threshold of 0.5. However, as discussed in Section 2.10 performance at a fixed threshold combines two distinct properties: the ability of a model to rank students according to dropout risk, and the choice of the operating point at which students are classified as being at risk. To separate these effects, Table 11 reports the PR-AUC, lift over the prevalence of the *dropout* class, ROC-AUC, and Brier score for every training configuration and model.

The threshold-free results qualify several conclusions suggested by the fixed-threshold metrics. First, the high *dropout* recall observed for CES is strongly influenced by its high dropout prevalence. Although PR-AUC reaches 0.703–0.729, the positive class constitutes approximately 60% of the test partition. This yields lifts of only 1.18–1.22, compared with 1.44–1.61 for the complete dataset. Information Technology shows an even smaller improve-

ment over its prevalence baseline, with lifts of 1.03–1.10. Thus, the relatively high *dropout* recall in these settings should not be interpreted as correspondingly strong discrimination.

Table 11. Threshold-free discrimination and calibration on the held-out test partition of each training configuration.

Scenario	Model	Prevalence	PR-AUC	Lift	ROC-AUC	Brier
Scenario 1: Complete dataset	Decision Tree	0.36	0.524	1.44	0.671	0.215
	Random Forest		0.555	1.52	0.691	0.207
	XGBoost		0.586	1.61	0.708	0.202
Scenario 2: Complete generations	Decision Tree	0.36	0.531	1.48	0.680	0.213
	Random Forest		0.577	1.61	0.711	0.200
	XGBoost		0.602	1.68	0.730	0.195
Scenario 3: Temporal split	Decision Tree	0.43	0.542	1.25	0.620	0.266
	Random Forest		0.575	1.33	0.654	0.255
	XGBoost		0.604	1.39	0.679	0.248
Scenario 4: Single center (CTC)	Decision Tree	0.37	0.508	1.35	0.652	0.225
	Random Forest		0.550	1.47	0.682	0.213
	XGBoost		0.559	1.49	0.683	0.211
Scenario 4: Single center (CES)	Decision Tree	0.60	0.703	1.18	0.625	0.231
	Random Forest		0.729	1.22	0.650	0.225
	XGBoost		0.726	1.22	0.647	0.226
Scenario 5: Computer Science	Decision Tree	0.43	0.584	1.35	0.629	0.234
	Random Forest		0.611	1.42	0.639	0.230
	XGBoost		0.625	1.45	0.660	0.227
Scenario 5: Languages	Decision Tree	0.44	0.536	1.22	0.606	0.239
	Random Forest		0.582	1.33	0.645	0.230
	XGBoost		0.582	1.33	0.648	0.230
Scenario 5: Information Technology	Decision Tree	0.53	0.541	1.03	0.529	0.272
	Random Forest		0.572	1.09	0.530	0.256
	XGBoost		0.579	1.10	0.513	0.285
Oversampling	Decision Tree	0.36	0.516	1.42	0.660	0.231
	Random Forest		0.543	1.49	0.680	0.221
	XGBoost		0.575	1.58	0.698	0.217
Undersampling	Decision Tree	0.36	0.518	1.42	0.667	0.232
	Random Forest		0.549	1.51	0.684	0.226
	XGBoost		0.580	1.59	0.705	0.218

Second, the sharp decrease in *dropout* recall under temporal evaluation is primarily an operating point and calibration problem rather than a comparable collapse in ranking ability. For XGBoost, recall at the default threshold decreases from 0.37 on the complete dataset to 0.18 under the temporal split, whereas ROC-AUC decreases more moderately from 0.708 to 0.679 and PR-AUC lift from 1.61 to 1.39. This distinction is also evident in the reliability diagrams shown in Figure 6. On the complete dataset, predicted and observed risks are closely aligned; for XGBoost, the mean difference between observed and predicted risk is -0.002. Under the temporal split, observed dropout rates systematically exceed predicted probabilities, with mean differences of $+0.155$ for XGBoost, $+0.163$ for Random Forest, and $+0.167$ for Decision Tree. Brier scores likewise increase from 0.202–0.215 on the complete dataset to 0.248–0.266 under temporal evaluation.

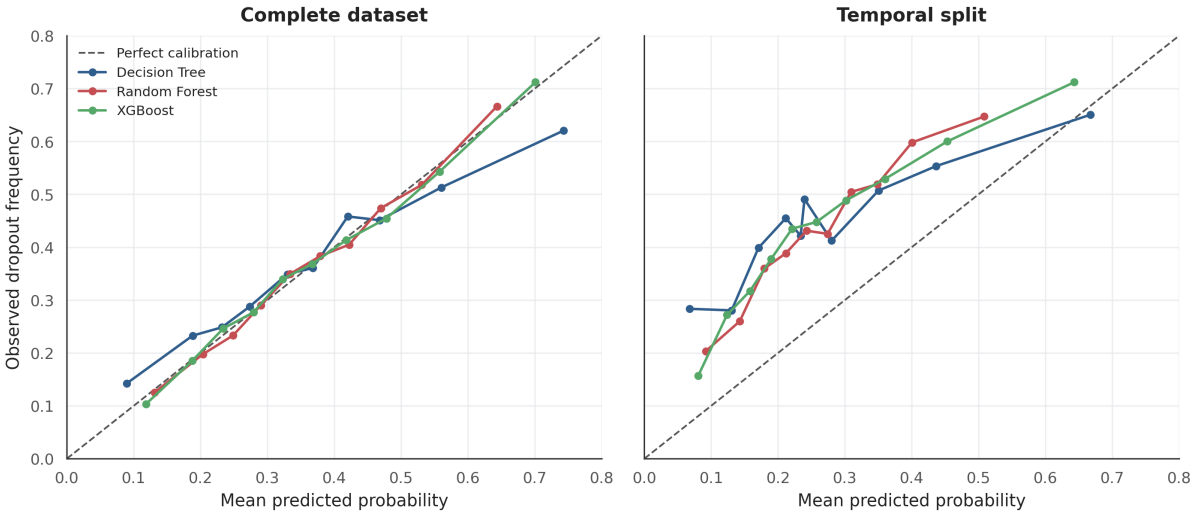


Figure 6. Reliability diagrams on the held-out test partitions.

The rolling-origin analysis in Scenario 3 provides complementary evidence. ROC-AUC varies little across the seven temporal cuts, with a standard deviation no greater than 0.011 for any model, whereas recall at the fixed threshold is both low and more variable. When the operating point is adjusted using the prevalence-matched threshold described in Section 2.5, mean *dropout* recall increases to 0.43, 0.43, and 0.45 for Decision Tree, Random Forest, and XGBoost, respectively, while its standard deviation falls to approximately 0.02–0.03. Mean PR-AUC lift across the rolling origins remains 1.25, 1.35, and 1.41, respectively. Together, these results indicate that temporal drift affects the probability scale, and hence performance at a fixed threshold, more strongly than it affects the models' ability to rank students by risk.

This distinction has an important implication for temporal deployment. A threshold selected from historical cohorts should not be assumed to remain appropriate for later cohorts. Rather than retaining the default threshold of 0.5, risk calibration and the operating threshold should be reassessed using recent outcome data as they become available. The attribution and sensitivity analyses in Section 3.7.1 and Appendix C examine specific changes in predictor distributions across the temporal split.

Third, resampling substantially changes fixed-threshold recall without improving threshold-free discrimination. For XGBoost, PR-AUC is 0.586 without resampling, compared with 0.575 after oversampling and 0.580 after undersampling; the same pattern occurs for Decision Tree and Random Forest. Therefore, the increase in *dropout* recall reported in Section 3.4 primarily reflects a shift in operating point rather than an improvement in ranking quality. Figure 7 illustrates this directly for XGBoost: the undersampled model obtains a recall of 0.64 at precision 0.51 using the default threshold, whereas the model without resampling reaches a similar trade-off by reducing its threshold to approximately 0.35, yielding recall 0.68 at precision 0.50.

These results favor treating the decision threshold as an explicit operational parameter rather than relying on class rebalancing to determine the effective operating point implicitly. An institution can then choose a threshold according to its capacity for intervention and its tolerance for false positives without changing the model or the training distribution.

As an illustrative decision rule, consider limiting the false positive rate to 0.20. For XGBoost on the complete dataset, this constraint yields a threshold of 0.45. At the default threshold of 0.5, the model flags 2922 of 13,364 students (21.9%), correctly identifying 1808 of the 4867 students who subsequently drop out, for a recall of 0.37 and 1114 false positives. At a threshold of 0.45, it flags 3919 students (29.3%), correctly identifies 2243 dropout cases,

and reaches a recall of 0.46 with 1676 false positives. Thus, accepting 562 additional false positives identifies 435 additional students who subsequently drop out.

Whether this trade-off is appropriate depends on the capacity, cost, and expected benefit of the intervention available to the institution. These quantities are not observed in the present dataset, so no universal operating threshold can be recommended. Instead, the analysis demonstrates how the threshold can be selected under an explicit institutional constraint and clarifies the resulting trade-off between coverage and false positives.

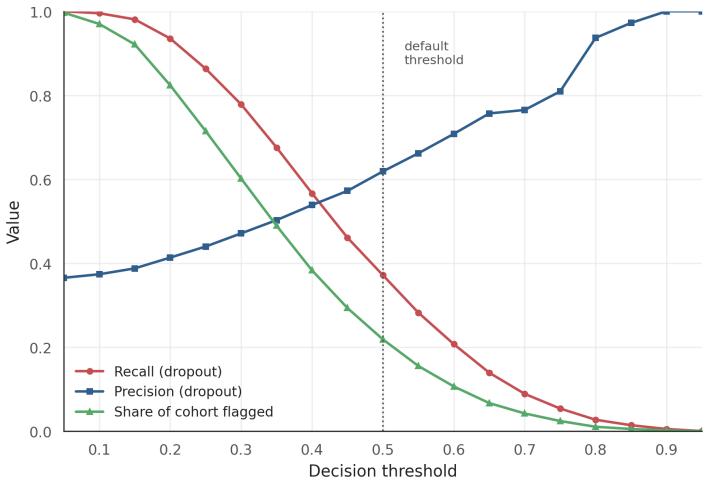


Figure 7. Effect of the decision threshold on the share of the cohort flagged for follow-up, *dropout* recall, and *dropout* precision for XGBoost on the complete dataset. The dotted line marks the default threshold of 0.5.

3.6. Comparison with Reference Baselines

The preceding scenarios compare three tree-based models, but do not establish whether their performance is specific to this modeling family. To provide broader reference points, the regularized logistic regression and the feed-forward neural network described in Section 2.6 were evaluated on the two partitions supporting the main institution-wide conclusions of the study, namely, the complete dataset and the temporal split. Table 12 compares their performance with that of the three tree-based models.

Table 12. Comparison of the tree-based models with the regularized logistic regression and the feed-forward neural network. Precision, recall, and F1-score refer to the *dropout* class.

Scenario	Model	Precision	Recall	F1-Score	PR-AUC
Complete dataset	Decision Tree	0.57	0.31	0.40	0.52
	Random Forest	0.60	0.32	0.42	0.56
	XGBoost	0.62	0.37	0.46	0.59
	Logistic Regression	0.59	0.33	0.43	0.56
	Neural Network	0.59	0.43	0.49	0.57
Temporal split	Decision Tree	0.65	0.13	0.22	0.54
	Random Forest	0.69	0.07	0.13	0.58
	XGBoost	0.70	0.18	0.28	0.60
	Logistic Regression	0.71	0.14	0.24	0.60
	Neural Network	0.72	0.17	0.28	0.61

On the complete dataset, the differences in threshold-free discrimination across the five models are relatively small. PR-AUC ranges from 0.524 for Decision Tree to 0.586 for XGBoost, corresponding to lifts of 1.44 and 1.61, respectively, over the dropout prevalence of 0.364. Logistic regression reaches a PR-AUC of 0.557, essentially matching Random Forest

(0.555) and approaching XGBoost despite its substantially simpler linear formulation. This result indicates that much of the predictive signal captured by the tree-based ensembles is also accessible to a regularized linear model using the same enrollment-time information.

The comparison under temporal evaluation reinforces this interpretation. Logistic regression obtains a PR-AUC of 0.601, compared with 0.604 for XGBoost, while the neural network reaches 0.610. At the default threshold, however, *dropout* recall is low for all five models, ranging from 0.07 to 0.18. Thus, the temporal degradation identified in Sections 3.3.3 and 3.5 is not specific to the tree-based models: linear, ensemble, and neural models all exhibit poor *dropout* detection at the default threshold when evaluated on more recent cohorts.

The neural network provides a further illustration of the distinction between fixed-threshold performance and threshold-free discrimination discussed in Section 3.5. On the complete dataset, it achieves the highest *dropout* recall (0.43) and F1-score (0.49), compared with 0.37 and 0.46 for XGBoost. However, its PR-AUC is lower (0.570 versus 0.586), indicating that its higher recall at the default threshold does not correspond to better ranking performance across thresholds. XGBoost provides the strongest threshold-free discrimination on the complete dataset, whereas the neural network operates at a more recall-oriented point at the default threshold.

These comparisons indicate that the modest predictive performance observed in this study is not explained by the choice of tree-based algorithms. Models spanning linear, tree-based ensemble, and neural formulations achieve similar threshold-free discrimination when trained on the same enrollment-time predictors, and the simpler logistic regression is competitive with the more complex models in both institution-wide evaluations. Combined with the generally weak predictor-outcome associations reported in Section 3.2, these results suggest that predictive performance is constrained mainly by the information available at enrollment. More complex algorithms alone may therefore provide limited gains unless they are accompanied by additional informative predictors.

3.7. Model Interpretation

To examine which enrollment-time variables contribute most strongly to model predictions, we analyzed both model-specific feature importance and SHAP values. The analyses focus on the complete-dataset and temporal-split scenarios, which provide complementary views of model behavior under random and temporally separated evaluation settings.

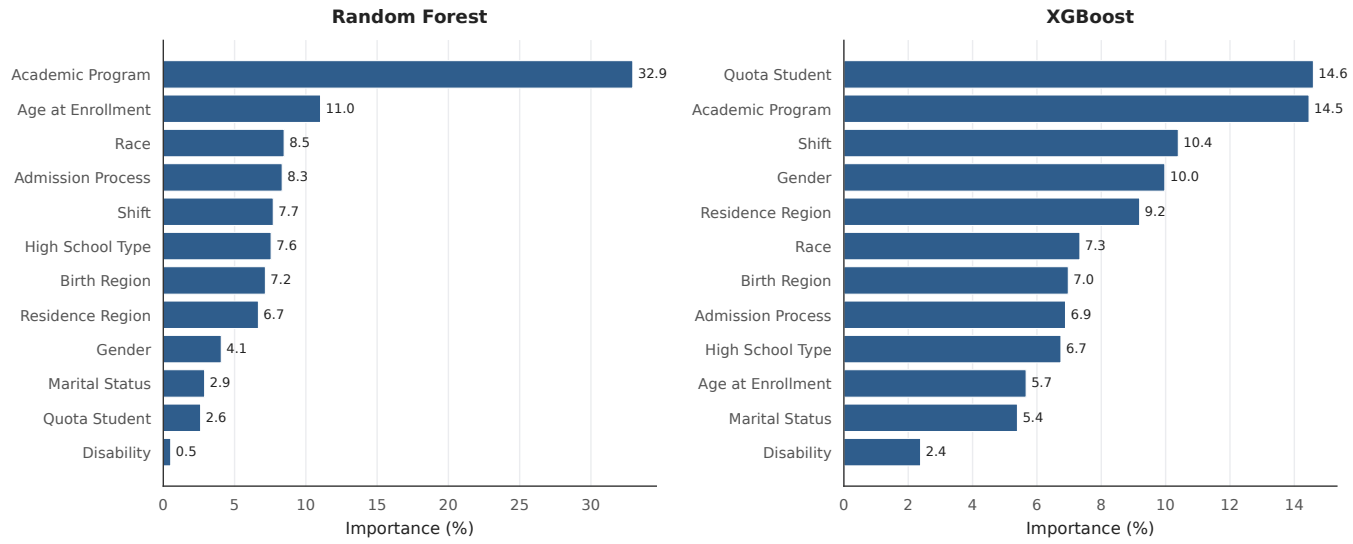
3.7.1. Global Feature Importance and SHAP Analysis

Figures 8 and 9 present the model-specific feature importance scores for Random Forest and XGBoost on the complete dataset and temporal split, respectively.

On the complete dataset, the two models differ substantially in how they distribute importance. For Random Forest, undergraduate program is the dominant predictor (32.9%), followed at a distance by age at enrollment, race, and admission process. XGBoost distributes importance more evenly, assigning its highest values to quota-student status (14.6%) and undergraduate program (14.5%), followed by study shift, gender, and residence region. Despite these differences, both models assign substantial importance to variables describing institutional context, admission conditions, and student characteristics.

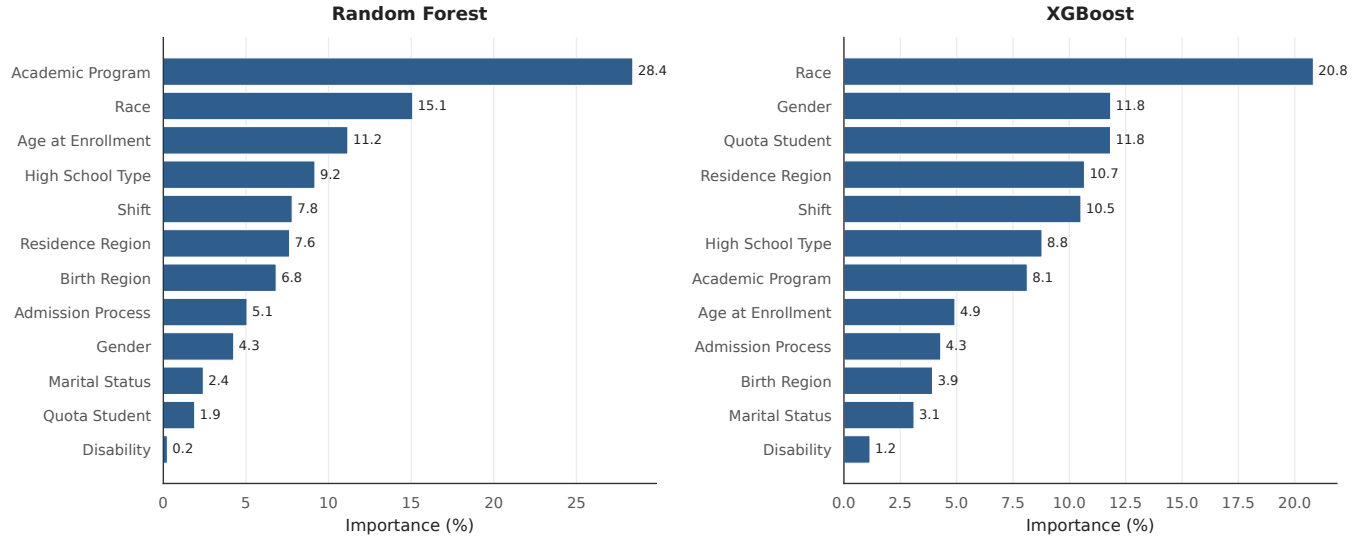
Under the temporal split, the relative importance of some predictors changes. Undergraduate program remains the most important variable for Random Forest (28.4%), while race rises to second place (15.1%). For XGBoost, race becomes the highest-ranked variable (20.8%), followed by gender, quota-student status, and residence region, while the importance assigned to undergraduate program decreases to 8.1%. These changes indicate

that model-specific importance scores are sensitive to the cohort structure under which the models are trained and evaluated.



(**a**) Random Forest (**b**) XGBoost

Figure 8. Feature importance for models trained on the complete dataset.



(**a**) Random Forest (**b**) XGBoost

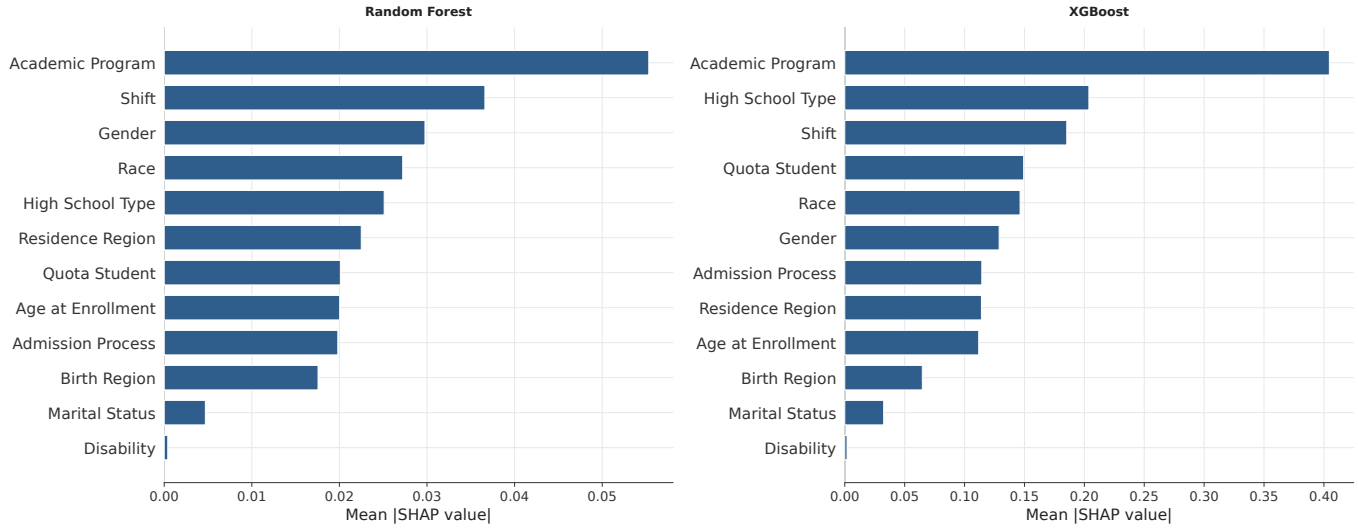
Figure 9. Feature importance for models trained on older generations and tested on more recent ones.

Importantly, model-specific importance measures require caution. The scores used here are based on impurity reduction for Random Forest and gain for XGBoost; such measures may favor predictors with many possible values, and may distribute importance among related predictors in ways that complicate interpretation. Both issues are relevant to this dataset: undergraduate program has 45 categories, and Section 3.2 identifies substantial associations among predictors, including undergraduate program with study shift ($V = 0.68$) and residence region with birth region ($V = 0.58$).

Therefore, we additionally complemented these measures with SHAP values [33]. SHAP attributes individual predictions to their contributing features, and was computed here on the held-out test partitions. Figures 10 and 11 report the mean absolute SHAP value

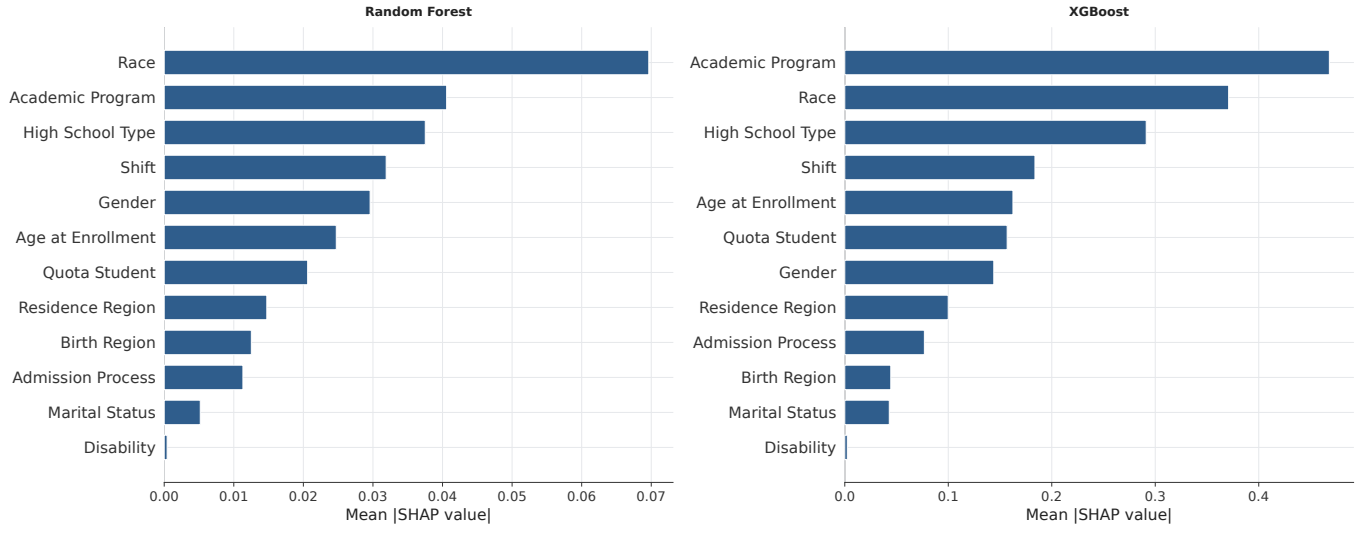
for each feature. For comparison across models, the values discussed below are expressed as percentages of the total mean absolute SHAP value across features.

The SHAP analysis confirms the broad importance of undergraduate program while modifying the relative weights suggested by the model-specific measures. On the complete dataset, undergraduate program is the leading feature for both models, accounting for 20% of total mean absolute SHAP importance for Random Forest and 24% for XGBoost, followed by study shift, race, gender, and high school type. For XGBoost, this differs notably from the gain-based ranking, in which quota-student status (14.6%) and undergraduate program (14.5%) were essentially tied; under SHAP, quota-student status accounts for 9% and undergraduate program for 24%. Conversely, the dominance of undergraduate program in the Random Forest impurity-based scores decreases from 32.9% to 20% under SHAP. Thus, the two approaches broadly agree on which variables contribute to prediction, but differ in how strongly importance is concentrated among them.



(**a**) Random Forest (**b**) XGBoost

Figure 10. Mean absolute SHAP value per feature for models trained on the complete dataset, computed on the held-out test partition.



(**a**) Random Forest (**b**) XGBoost

Figure 11. Mean absolute SHAP value per feature for models trained on older generations and tested on more recent ones, computed on the held-out test partition.

Despite changes in individual importance values, the overall ranking of predictors remains relatively stable across cohort definitions. The Spearman rank correlation between the complete-dataset and temporal-split importance vectors is $+0.86$ for Decision Tree, $+0.89$ for Random Forest, and $+0.87$ for XGBoost. Race is a notable exception to this overall stability; its share of total importance increases across scenarios involving progressively more temporally restricted cohorts, from 10%, 10%, and 9% on the complete dataset (Decision Tree, Random Forest, and XGBoost, respectively) to 16%, 15%, and 12% for complete generations and to 24%, 23%, and 18% under the temporal split configuration. The consistency of this pattern across all three models suggests that it is not specific to a single algorithm.

This pattern coincides with a substantial change in the distribution of the recorded race categories across cohorts. Among the cohorts used for training in the temporal split, 51.6% of students have their race recorded as *not declared*, compared with only 3.8% in the test cohorts. Over the same partitions, the proportion recorded as *white* increases from 37.4% to 72.8% and the proportion recorded as *mixed* increases from 6.4% to 15.1%. The *not declared* category also has a higher dropout rate on both sides of the split: 42.2% in training and 59.2% in testing, compared with 23.6% and 42.0%, respectively, for the *white* category. Consequently, a category that is both frequent and informative in the historical cohorts becomes uncommon in the more recent cohorts.

A related temporal change occurs in quota admissions, where prevalence increases from 3.7% in the training cohorts to 14.0% in the test cohorts. Quota status is also associated with race ($V = 0.23$; Section 3.2). Together, these changes illustrate why demographic and administrative predictors require particular caution in longitudinal institutional datasets: their apparent predictive importance may reflect changes not only in the student population but in institutional policies and recording practices. The substantial change in the *not declared* category motivates the sensitivity analysis reported in Appendix C.

Overall, these results show that feature importance depends on both the model and the temporal context in which it is evaluated. Variables describing academic structure and student admission profiles consistently contribute to prediction, but their relative importance can change across cohorts. Therefore, importance rankings should be interpreted as predictive associations rather than causal effects; particularly for longitudinal institutional data, they should also be examined together with changes in data collection and institutional context.

3.7.2. Program Structure and Individual Risk

The prominence of undergraduate program in the importance analyses raises a further question around the extent to which the models distinguish students within the same program rather than ranking programs primarily according to their historical dropout rates. Table 13 examines this question on the complete dataset by comparing three complementary settings: prediction using only the program-level dropout rate, models fitted without undergraduate program, and discrimination among students within the same program.

Using only the dropout rate of each student's undergraduate program as estimated from the training partition produces a PR-AUC lift of 1.382 and a ROC-AUC of 0.644. Thus, program structure alone provides substantial predictive information; the complete models attain lifts of 1.438 for Decision Tree, 1.525 for Random Forest, and 1.610 for XGBoost. The relatively small difference between the program-only baseline and the fitted models confirms that part of their predictive performance reflects differences in dropout prevalence across programs.

The within-program ROC-AUC provides a complementary test, since students are compared only with others enrolled in the same program. In this test, undergraduate program

is constant within each comparison, and cannot determine the ordering. The program-base-rate rule has a within-program ROC-AUC of 0.500 by construction, whereas Decision Tree, Random Forest, and XGBoost achieve 0.613, 0.638, and 0.663, respectively. Thus, the fitted models retain substantial ability to rank students within the same undergraduate program, indicating that their discrimination is not explained by program-level differences alone.

Table 13. Contribution of undergraduate program to predictive performance on the complete dataset (Scenario 1). Lift is PR-AUC divided by the prevalence of the *dropout* class (0.364). The within-program ROC-AUC ranks students only against others in the same program, so a value of 0.500 indicates a model carrying no individual information.

	Lift			
Predictors	**Without Program**	**All Predictors**	**ROC-AUC**	**Within-Program ROC-AUC**
Program base rate only	—	1.382	0.644	0.500
Decision Tree	1.334	1.438	0.671	0.613
Random Forest	1.402	1.525	0.691	0.638
XGBoost	1.449	1.610	0.708	0.663

This conclusion is supported by the models refitted without undergraduate program. Their PR-AUC lifts remain at 1.334 for Decision Tree, 1.402 for Random Forest, and 1.449 for XGBoost. Therefore, for Random Forest and XGBoost, the remaining enrollment-time predictors alone provide discrimination comparable to or greater than that obtained from program-level dropout rates alone. At the same time, the program-level and individual-level information are not independent; as shown in Section 3.2, undergraduate program is strongly associated with study shift ($V = 0.68$) and moderately associated with gender ($V = 0.42$).

The same distinction between program-level prevalence and individual discrimination is relevant when performance is examined across all undergraduate programs. Figure 12 reports XGBoost performance for the 45 programs represented in the complete-dataset test partition. The two panels use complementary measures because recall at a fixed threshold is affected by the prevalence of dropout within each program, whereas within-program ROC-AUC evaluates the ranking of students independently of differences in program-level prevalence.

At the default threshold of 0.5, program-level *dropout* recall is strongly associated with the program's own dropout rate (Spearman $\rho = +0.85$). Thus, programs with higher dropout prevalence tend to show higher recall even when this does not correspond to better discrimination among their students. In contrast, the association between dropout prevalence and within-program ROC-AUC is weak ($\rho = -0.18$). This distinction is consistent with the prevalence effects identified in Section 3.5, and shows why recall alone should not be used to compare predictive performance across programs with substantially different dropout rates.

Within-program ROC-AUC ranges from 0.513 to 0.805, with a median of 0.649. None of the 45 programs falls below 0.500, indicating that the model retains some ability to discriminate among students within each program, although the strength of this discrimination varies substantially. Medicine and the Sciences teaching degree illustrate the difference between fixed-threshold recall and ranking performance: their within-program ROC-AUC values are nearly identical (0.769 and 0.768), whereas their *dropout* recall values at the 0.5 threshold are 0.04 and 1.00, respectively, alongside dropout prevalence scores of 0.13 and 0.67. To a substantial extent, the large difference in recall reflects the different operating conditions created by program-level prevalence rather than a corresponding difference in ranking quality.

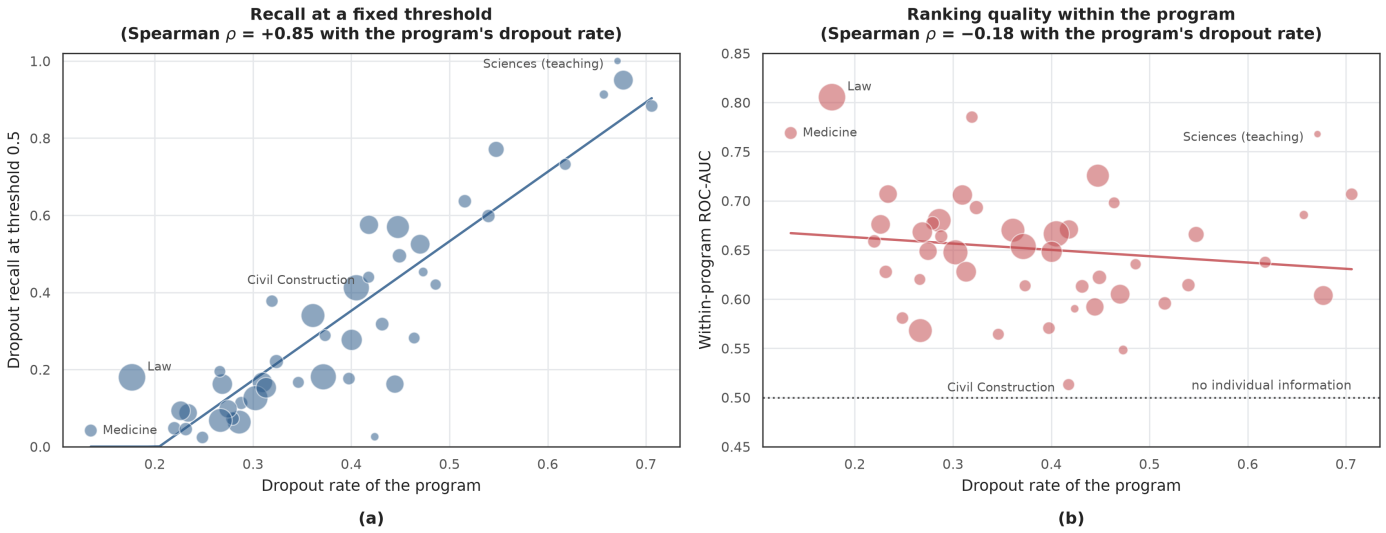


Figure 12. Performance of XGBoost within each of the 45 undergraduate programs of the complete-dataset test partition against the program's own *dropout* rate. Point area is proportional to the number of students. (**a**) *Dropout* recall at the fixed threshold of 0.5. (**b**) ROC-AUC computed within the program, which ranks a program's students only against each other and as such is independent of its *dropout* rate; 0.500 marks a model carrying no individual information.

These findings clarify how program-related predictive information should be interpreted. Undergraduate program captures an important structural component of predicted dropout risk, but the models also extract information that distinguishes students within the same program. Neither result establishes a causal effect of program on dropout. Program membership may reflect differences in student composition, study conditions, curricular and institutional structures, and historical dropout prevalence, some of which are not directly represented by the available enrollment-time predictors.

More generally, most enrollment-time attributes used by the models, including undergraduate program, age at enrollment, race, birth region, high school type, and admission route, describe conditions that are already established when the student enters the university; thus, they should be interpreted as risk markers rather than intervention targets. Their practical value lies in identifying students who may benefit from support, not in implying that the attributes themselves should or could be modified.

The prominence of program and other institutional or demographic attributes also raises a question that feature attribution and predictive performance alone cannot answer: whether the model behaves systematically differently across student groups. Section 3.8 examines this question directly.

3.8. Fairness and Institutional Bias

Table 14 reports the performance of the XGBoost model on the complete-dataset test partition, broken down by student group. Within each group with at least 100 students in the test partition, the mean predicted dropout probability is close to the observed dropout rate, differing by at most 0.04, indicating that the model is well calibrated for these groups. The groups differ in the errors made by the model at the default threshold. The difference between the highest and lowest group recall is 0.33 for race, 0.19 for age at enrollment, 0.16 for gender, and 0.14 for quota admission, with parallel differences in the false positive rate. High school type is the exception, with recall and false positive rate almost identical for public and private students, differing by 0.02 and 0.01, respectively.

These differences follow those in observed dropout rates across groups. Groups with a higher dropout rate, namely, male students, students who entered at age 25 or older, quota

students, and students whose race is *not declared* are assigned higher predicted probabilities and consequently are flagged more often, which raises both their recall and their false positive rate. Because the mean predicted dropout probabilities closely track the observed dropout rates across groups, a single fixed threshold produces different error rates when the underlying dropout rates differ; this trade-off between calibration and equal error rates across groups is a known property of group-level fairness [34–36]. The pattern is consistent with this explanation: high school type, the one attribute for which the dropout rate is similar across groups, shows almost no difference in error rates.

Table 14. Fairness evaluation of the XGBoost model on the complete-dataset test partition ($n = 13{,}364$), at the default probability threshold of 0.5. For each group: number of students (n), observed dropout rate (base rate), proportion flagged as *dropout* (flag rate), recall for the *dropout* class (true positive rate), false positive rate (FPR), mean predicted dropout probability, and ROC-AUC. The max–min gap is the difference between the largest and smallest group values of recall, FPR, and AUC.

Group	*n*	Base Rate	Flag Rate	Recall	FPR	Mean Pred.	AUC
Gender							
Female	6927	0.33	0.15	0.29	0.09	0.33	0.70
Male	6437	0.40	0.29	0.45	0.18	0.40	0.71
Max-min gap				0.16	0.09		0.01
Race							
Not declared	3527	0.43	0.37	0.55	0.23	0.43	0.73
White	7314	0.33	0.15	0.27	0.09	0.33	0.70
Mixed (pardo)	1605	0.39	0.23	0.35	0.15	0.38	0.67
Black	291	0.34	0.22	0.40	0.13	0.38	0.70
Asian	595	0.32	0.12	0.22	0.08	0.32	0.67
Indigenous †	32	0.59	0.25	0.37	0.08	0.41	0.67
Max-min gap				0.33	0.16		0.06
Age at enrollment							
≤18	7406	0.35	0.20	0.35	0.12	0.36	0.71
19–24	4650	0.36	0.20	0.35	0.12	0.35	0.71
25+	1307	0.47	0.40	0.53	0.29	0.46	0.67
Max-min gap				0.19	0.17		0.04
Quota admission							
No	12,244	0.36	0.21	0.36	0.12	0.36	0.71
Yes	1120	0.43	0.36	0.50	0.26	0.44	0.66
Max-min gap				0.14	0.14		0.05
High school type							
Public	8035	0.35	0.21	0.36	0.13	0.36	0.71
Private	5320	0.38	0.23	0.38	0.14	0.37	0.71
Abroad †	9	0.44	0.56	0.75	0.40	0.50	0.60
Max-min gap				0.02	0.01		0.00

† Fewer than 100 students in the test partition; reported but excluded from the max-min gap.

Two points deserve emphasis: first, the largest apparent disparity, that for race, is driven by the *not declared* category rather than by differences among the declared racial groups. As discussed in Section 3.7.1, this category reflects how the information was recorded rather than a student characteristic, and its high dropout rate makes it the most heavily flagged group. Second, quota students are the only group for which discrimination itself is weaker: their ROC-AUC of 0.66 is lower than the 0.71 of non-quota students, so for this group the model ranks students slightly less well.

3.9. Comparison with Prior Work

Direct numerical comparison with previously published dropout models requires caution because studies differ in institution, outcome definition, predictor availability, class prevalence, and validation design. We interpret the results in relation to three aspects of the literature that are directly relevant to this study: prediction using enrollment-time information, prediction using academic trajectory data, and methodological choices concerning class imbalance and temporal validation.

Among studies restricted to information available at or before enrollment, a particularly relevant comparison is the early-warning system of Carballo-Mendívil et al. [19]. Using XGBoost and enrollment-time data from approximately 40,000 first-year students at a Mexican public university, that study reports a ROC-AUC of 0.690. In our experiments, XGBoost achieves a ROC-AUC of 0.708 on the complete dataset and 0.679 under temporal evaluation. Despite substantial differences between the two institutional settings, these values indicate a similar level of threshold-free discrimination. The same study reports 88% sensitivity for the *dropout* class while classifying approximately 59% of incoming students as high risk. Recall values reported by different studies must therefore be read against their operating points: as shown in Section 3.5, higher recall can be obtained by flagging a larger proportion of students without any improvement in the model's ability to rank them by risk. Studies using pre-university achievement data, such as Nagy and Molontay [37], likewise report predictive performance within a comparable range. Together with the reference-baseline results in Section 3.6, these comparisons suggest that the moderate discrimination observed here is not specific to the tree-based algorithms evaluated in this study.

Studies that incorporate information accumulated after enrollment often report higher predictive performance. For example, Cho et al. [14] report an F1-score of 0.84 for the *dropout* class using accumulated academic records from approximately 20,000 students, whereas the highest F1-score among our models on the complete dataset at the default threshold is 0.49 for the neural-network reference model and 0.46 among the three primary tree-based models. These values are not directly comparable because the studies differ in population, class distribution, predictors, and experimental design, but the contrast is consistent with evidence that academic trajectory information can substantially improve dropout prediction. Both Fernández-García et al. [22] and Berens et al. [38], for example, report improved predictive performance as information from successive stages of students' academic trajectories becomes available. Our results reflect the other side of this trade-off: restricting predictors to enrollment-time information limits what the model can learn, but allows predictions before academic performance records accumulate. This timing is particularly relevant in our dataset because dropout events are concentrated early in students' trajectories, with the largest number occurring within the first six months after enrollment (Figure 2).

Our findings also provide additional context for two methodological practices commonly used in dropout prediction. First, studies comparing class-balancing techniques often select a preferred resampling method according to classification metrics obtained under random train–test partitions [15,39]. In our experiments, both oversampling and undersampling substantially increase *dropout* recall at the default threshold, but their threshold-free discrimination remains similar to and slightly below that obtained without resampling (Sections 3.4 and 3.5). The paired comparisons further show that differences between the two resampling strategies are small relative to the effect of rebalancing itself. These results suggest that improvements in fixed-threshold recall after resampling should not automatically be interpreted as improvements in the underlying discrimination of the model.

Second, temporal evaluation produces a substantially different picture from random partitioning. Previous work has documented temporal and between-group variability in dropout prediction [16]. Our temporal and rolling-origin experiments extend this observation by showing that discrimination remains more stable than recall at the default threshold, while calibration changes substantially across cohorts. In particular, the models systematically underestimate the dropout risk observed in the more recent cohorts (Section 3.5). This finding strengthens the case for temporally ordered evaluation when the intended application is prediction for future cohorts and indicates that model deployment should account explicitly for changes in class prevalence and probability calibration over time.

4. Conclusions

This study investigated academic dropout prediction using large-scale institutional data spanning 20 years and restricted to attributes available at enrollment. Five training scenarios were evaluated to examine how cohort definition, temporal structure, institutional context, and class distribution affect predictive performance. The findings discussed throughout this paper can be summarized into four main conclusions.

First, model performance depends substantially on the data configuration under which it is evaluated. XGBoost generally achieved the strongest performance in the larger institution-wide scenarios, but the differences among model families were modest when threshold-free discrimination was considered. The reference baselines also showed that regularized logistic regression and a feed-forward neural network achieved performance close to the tree-based models, while repeated-split experiments at the program level revealed substantial uncertainty in model rankings when sample sizes were small. These findings suggest that the choice of algorithm matters less than the amount, composition, and temporal structure of the data available for prediction.

Second, temporal evaluation revealed a substantially more difficult prediction setting than random train-test partitioning. When models trained on older generations were evaluated on more recent ones, *dropout* recall at the default threshold decreased sharply. Threshold-free discrimination deteriorated far less, while the reliability analysis showed systematic underestimation of dropout risk in the newer cohorts. The rolling-origin experiments confirmed that this pattern persists across multiple temporal cut points and that adjusting the operating threshold can recover a substantial portion of *dropout* recall without changing the underlying model. These results show why evaluation should be temporally ordered and why calibration and threshold selection should be treated as explicit components of model deployment rather than left at a fixed default threshold.

Third, the apparent benefit of class-imbalance handling was primarily associated with the operating point rather than with improved discrimination. Both oversampling and undersampling substantially increased *dropout* recall at the default threshold, but neither improved PR-AUC relative to the original training distribution. A comparable precision-recall trade-off could be obtained from the unresampled model by lowering its decision threshold. Thus, when the objective is to identify a larger proportion of at-risk students, explicit threshold selection may provide a more transparent and adjustable mechanism than modifying the training distribution.

Fourth, model interpretation showed that the undergraduate program was the most influential predictor of dropout, while admission-related and demographic attributes also contributed to prediction. At the same time, predictor importance was not fully stable across cohorts. The temporal SHAP analysis revealed changes in the attributed importance of several variables, some of which coincided with substantial shifts in their recorded distributions, particularly for undeclared values. The sensitivity analysis showed that alternative treatments of these values did not eliminate the degradation observed under

temporal evaluation, indicating that this degradation cannot be attributed solely to changes in the recording of undeclared information. More broadly, the moderate discrimination achieved across tree-based, linear, and neural models, together with the generally weak individual associations between enrollment-time predictors and dropout, indicates the limitations of prediction based exclusively on information available at enrollment.

Finally, these findings show that random splits and fixed-threshold classification metrics alone give an incomplete picture of how dropout prediction models would behave in practice [16,21]. Robust assessment should consider temporal generalization, class prevalence, probability calibration, the stability of predictor contributions across cohorts, and the operational consequences of threshold selection. Future work should investigate whether incorporating additional information available early in students' trajectories can improve discrimination while preserving the possibility of timely intervention, and should evaluate the generalizability of these findings across institutions with different student populations, academic structures, and data-recording practices.

4.1. Limitations

This study has three main limitations. First, the exclusive use of static enrollment-time variables restricts the ability to capture dynamic aspects of students' academic experience that are frequently associated with dropout in the literature [4,13]. The comparison with the reference baselines in Section 3.6 shows that models of substantially different complexity can achieve relatively similar levels of threshold-free discrimination, suggesting that the moderate predictive performance observed here is related at least in part to the information available in the enrollment-time attributes rather than to the choice of model family alone. This interpretation is also consistent with the theoretical accounts discussed in the Introduction. Constructs regarded as proximal determinants of withdrawal, such as academic and social integration [3] and the self-efficacy beliefs emphasized by social cognitive theory [6], are dynamic quantities that administrative enrollment records do not directly operationalize; therefore, longitudinal academic or behavioral information could improve prediction, although incorporating such information would necessarily move the prediction point further into the student's academic trajectory.

Second, the analysis was conducted at a single institution, which limits the generalizability of the findings to educational contexts with different program structures, admission policies, and student populations. Several predictors examined here, such as quota-based admission, self-declared race, and high school type, are closely related to the Brazilian higher education context. In addition, UEM is a tuition-free public university with competitive admission and on-campus instruction, a setting that differs substantially from private institutions and distance education programs. The specific performance values and predictor importance patterns reported here should not be assumed to transfer directly to other institutions.

Although the temporal evaluation framework itself is not specific to UEM, replicating it requires sufficiently large and long-duration institutional records. Because a generation is considered complete only when its admission year plus the maximum allowed program duration falls within the observation window, the longitudinal span of the data is a fundamental constraint. At UEM, the maximum program duration reaches approximately eight years; an institution with a shorter history of digitized records may have few or no complete generations available for temporal evaluation, irrespective of its total number of students. The rolling-origin analysis also required sufficiently large partitions on both sides of each temporal cut. In this study, we retained cuts with at least approximately 3000 students in the test cohorts and with a training partition no smaller than the test partition. These values should be understood as practical criteria used to obtain informative

temporal comparisons in this dataset rather than as more general minimum sample size requirements. Similarly, program-level modeling should be interpreted cautiously for small programs. In the programs examined here, repeated-split confidence intervals overlapped substantially, indicating considerable uncertainty in the model rankings.

Finally, the binary target compresses heterogeneous student trajectories into a single outcome. Canceled enrollments account for 97.0% of the *dropout* class (Table 5), and the institutional records available for this study do not distinguish the different circumstances leading to cancellation. Consequently, voluntary withdrawal cannot be separated from other forms of enrollment cancellation without access to information that is not represented in the dataset. The target is also program-centered: a student who transfers to another program is classified as a dropout from the original program, whereas trajectories such as temporary interruption followed by re-enrollment or transfer to another institution may not be fully represented by a single institution's terminal-status records. Therefore, the predictions should be interpreted according to the operational definition of dropout adopted in this study rather than as distinguishing among the different behavioral or administrative processes that may lead to non-completion. A finer-grained analysis would require more detailed institutional status information.

4.2. Deployment Protocol

The temporal results have direct implications for deployment. In the experiments conducted here, the sharp decline in *dropout* recall across cohorts was accompanied by a much smaller deterioration in threshold-free discrimination and by substantial changes in probability calibration. This suggests that threshold adjustment and calibration monitoring should precede retraining when similar temporal changes are observed in practice. The following protocol translates these findings into operational steps.

A central constraint is that the outcome used to evaluate the model is not immediately observable. A student's final status may remain unresolved until the maximum allowed duration of the program has elapsed (Section 2.4), so the *dropout* recall achieved on a cohort scored at enrollment cannot be measured in time to guide interventions for that same cohort. Monitoring must combine signals available at the time of scoring with retrospective performance evaluation once outcomes become sufficiently mature.

The first operational decision concerns the classification threshold. Rather than automatically adopting the default value of 0.5, the threshold should reflect the institution's intervention capacity and the intended trade-off between identifying students at risk and generating false positives. The rolling-origin experiments illustrate the importance of this choice: using a prevalence-matched threshold substantially increased *dropout* recall and reduced its variability across temporal cuts relative to the fixed threshold of 0.5 (Table 9) without retraining the models or using labels from the test cohorts. In deployment, however, prevalence matching should be regarded as an experimental demonstration rather than a required operating rule. If student support services can follow (for example) 20% of an incoming cohort, then the threshold can instead be selected to flag approximately that proportion. The precision–recall and workload trade-offs illustrated in Figure 7 can provide the basis for making this choice.

Because outcome-based performance cannot be measured immediately, short-term monitoring should focus on quantities available when a new cohort is scored. These include changes in predictor distributions, the distribution of predicted risks, and the proportion of students exceeding the operational threshold. The present data illustrate why such monitoring is necessary. Across the temporal analyses, dropout prevalence changed between older and newer cohorts (Section 3.3.3), while the distribution of several predictors also changed substantially; for example, the proportion of *not declared* values

for race decreased sharply, and quota admissions became more frequent (Section 3.7.1). Such changes can be detected before outcomes are known, allowing them to serve as early indicators that the population being scored differs from the data used to train the model.

The share of students flagged for intervention provides an additional operational signal. Under the temporal split, for example, Random Forest flagged only 4.3% of the test cohort at the default threshold, despite a substantially higher dropout prevalence being observed subsequently (Section 3.3.3). In prospective use, the true dropout prevalence of the incoming cohort would not yet be known, so flagged share alone cannot establish that the model is miscalibrated. A large departure from the intervention capacity for which the threshold was selected or from historically plausible levels of predicted risk should instead trigger review of the score distribution, predictor drift, and operating threshold. Where sufficiently recent cohorts with mature outcomes are available, calibration can then be assessed directly using reliability analysis such as that shown in Figure 6. In our experiments, the mean discrepancy between observed and predicted risk was approximately zero on the complete-dataset partition, but was substantially larger under temporal evaluation (Section 3.5).

For retraining, rather than being triggered automatically by a fixed calendar, it should be considered when retrospective evaluation provides evidence that discrimination itself has deteriorated. Across the rolling-origin cuts, ROC-AUC varied relatively little, whereas performance at the fixed threshold varied substantially (Section 3.3.3). This distinction suggests two different responses: changes primarily affecting calibration or the operating point can first be addressed through threshold adjustment or recalibration, whereas a sustained decline in threshold-free discrimination indicates that the ranking learned from historical data may no longer transfer adequately to newer cohorts. Once outcomes are available, PR-AUC should be interpreted relative to the contemporaneous dropout prevalence using the PR-AUC lift, as defined in Section 2.10. A persistent decline in PR-AUC lift across mature cohorts would provide evidence for model refitting and subsequent validation.

A separate monitoring requirement applies when model explanations are used to support institutional decisions. The analyses in Section 3.7.1 and Appendix C show that changes in administrative recording practices can substantially alter attributed feature importance even when they do not account for the observed change in predictive performance. Therefore, predictor distributions and attribution patterns should be monitored alongside predictive metrics; in particular, categories such as *not declared* in our dataset should be interpreted as indicators of how institutional information was recorded rather than as intrinsic student characteristics. For such categories, changes in their prevalence or attributed importance should prompt examination of the underlying data generation process before any substantive interpretations are made.

Author Contributions: Conceptualization: R.B.M., L.G.C., L.d.O.T., T.E.C., Y.M.e.G.d.C. and V.D.F.; implementation of the data processing and ML pipelines: R.B.M., L.G.C., L.d.O.T., and V.D.F.; preparation of figures and tables: R.B.M., L.d.O.T., L.G.C., and V.D.F.; data analysis: R.B.M., L.G.C., L.d.O.T., T.E.C., Y.M.e.G.d.C. and V.D.F.; writing and revision of the manuscript: R.B.M., L.G.C., L.d.O.T., T.E.C., Y.M.e.G.d.C. and V.D.F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Coordination for the Improvement of Higher Education Personnel (CAPES)–Finance Code 001, and by the National Council for Scientific and Technological Development (CNPq), Grant No. 308574/2023-0.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee of the State University of Maringá (COPEP), Brazil (CAAE 89887025.3.0000.0104; approval date: 14 July 2025).

Data Availability Statement: The dataset used in this study is unavailable due to privacy and ethical restrictions.

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5.6 Sol, OpenAI) to assist with language editing and improving the clarity and structure of the text. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CES	Center of Exact Sciences
CTC	Center of Technology
DPC	Data Processing Center
DT	Decision Tree
LR	Logistic Regression
ML	Machine Learning
NN	Neural Network
PR-AUC	Area Under the Precision–Recall Curve
RF	Random Forest
ROC-AUC	Area Under the Receiver Operating Characteristic Curve
SHAP	SHapley Additive exPlanations
SMOTE-NC	Synthetic Minority Oversampling Technique for Nominal and Continuous features
UEM	State University of Maringá
XGBoost	eXtreme Gradient Boosting

Appendix A. Modeling Variables

This appendix describes the predictors and target variable used for model training. As detailed in Section 2.2, the predictor set was restricted to information available at or before enrollment in order to prevent information leakage.

Table A1. Variables used for model training. The final row identifies the target variable; all preceding variables are predictors.

Variable	Description	Categories
Program	Undergraduate program in which the student was enrolled.	45
Study shift	Study shift associated with the student's program, such as morning, afternoon, evening, or combined schedules.	6
Marital status	Student's marital status at enrollment.	7
Race	Student's self-declared race.	6
Admission process	Type of admission process through which the student entered the university.	10
Gender	Student's recorded gender.	2
High school type	Type of high school attended by the student before admission, such as public or private school.	7
Quota admission	Indicator of whether the student was admitted through a quota-based admission policy.	2
Age at enrollment	Student's age at the time of enrollment, grouped into categorical intervals.	12
Residence region	Geographic region associated with the student's residence during vacation periods.	4
Birth region	Geographic region associated with the student's place of birth.	4
Disability status	Indicator of whether the student had a registered disability.	2
Dropout status	Binary target variable indicating dropout or non-dropout status.	2

Note: Age at enrollment was discretized into the following 12 intervals (in years): $\leq$16, 17–18, 19–20, 21–22, 23–24, 25–26, 27–30, 31–40, 41–50, 51–60, 61–70, and $>$70.

Appendix B. Extra Scenarios

This appendix contains detailed results that support the main findings reported in Section 3.

Appendix B.1. Scenario 4: Academic Centers

Table A2 compares model performance for the Center of Technology (CTC) and the Center of Exact Sciences (CES), which represent contrasting class distributions as described in Section 2.5.

Table A2. Performance of the models for the Center of Technology (CTC) and the Center of Exact Sciences (CES) in Scenario 4.

Center	Model	Class	Precision	Recall	F1-Score	Accuracy
CTC	Decision Tree	Non-dropout	0.67	0.84	0.74	0.64
		Dropout	0.53	0.30	0.38	
	Random Forest	Non-dropout	0.67	0.89	0.76	0.65
		Dropout	0.59	0.25	0.36	
	XGBoost	Non-dropout	0.68	0.86	0.76	0.66
		Dropout	0.57	0.32	0.41	
CES	Decision Tree	Non-dropout	0.55	0.29	0.38	0.62
		Dropout	0.64	0.84	0.72	
	Random Forest	Non-dropout	0.58	0.25	0.35	0.63
		Dropout	0.63	0.88	0.74	
	XGBoost	Non-dropout	0.57	0.34	0.42	0.63
		Dropout	0.65	0.83	0.73	

The two centers show markedly different class-specific performance. At CTC, where *non-dropout* is the majority class, *dropout* recall ranges from 0.25 to 0.32, whereas *non-dropout* recall ranges from 0.84 to 0.89. At CES, where *dropout* is the majority class, this pattern is reversed: *dropout* recall increases to 0.83–0.88, whereas *non-dropout* recall decreases to 0.25–0.34. Despite this reversal, overall accuracy remains similar across the two centers, ranging from 0.62 to 0.66.

The reversal in class-specific recall follows the contrasting class distributions of the two centers, indicating that performance at the default decision threshold is strongly influenced by the local outcome distribution. Consequently, the substantially higher *dropout* recall observed at CES should not be interpreted as evidence of better discrimination by itself. This comparison is revisited in Section 3.5 using threshold-free metrics and calibration analysis.

Appendix B.2. Scenario 5: Individual Programs

Table A3 compares model performance for Computer Science, Languages, and Information Technology, selected to represent different program-level class distributions as described in Section 2.5.

The single-split results vary across programs. For Computer Science, XGBoost obtains the highest *dropout* recall and F1-score (0.49 and 0.55), whereas Random Forest obtains the highest *dropout* recall for Languages (0.41) and Information Technology (0.58), with XGBoost performing similarly in both programs. Performance is also more symmetric between classes in Information Technology than in Computer Science and Languages.

Because the program-level test sets contain only a few hundred students, these single-split rankings were examined through the repeated-split stability analysis described in Section 2.5. Table A4 reports mean *dropout* recall and F1-score and their 95% confidence intervals across 50 stratified splits.

Table A3. Performance of the models for the three undergraduate programs evaluated in Scenario 5.

Program	Model	Class	Precision	Recall	F1-Score	Accuracy
Computer Science	Decision Tree	Non-dropout	0.64	0.81	0.72	0.63
		Dropout	0.62	0.40	0.49	
	Random Forest	Non-dropout	0.63	0.80	0.71	0.62
		Dropout	0.59	0.38	0.46	
	XGBoost	Non-dropout	0.67	0.77	0.71	0.65
		Dropout	0.62	0.49	0.55	
Languages	Decision Tree	Non-dropout	0.59	0.88	0.71	0.59
		Dropout	0.60	0.23	0.34	
	Random Forest	Non-dropout	0.64	0.81	0.71	0.63
		Dropout	0.62	0.41	0.49	
	XGBoost	Non-dropout	0.63	0.81	0.71	0.63
		Dropout	0.61	0.40	0.48	
Information Technology	Decision Tree	Non-dropout	0.49	0.73	0.58	0.51
		Dropout	0.56	0.30	0.39	
	Random Forest	Non-dropout	0.47	0.42	0.45	0.51
		Dropout	0.53	0.58	0.55	
	XGBoost	Non-dropout	0.50	0.51	0.51	0.53
		Dropout	0.55	0.54	0.55	

Table A4. Stability of *dropout* recall and F1-score across 50 repeated stratified splits within each program. Values are means over repeats, with 95% confidence intervals in brackets.

Program	Model	Recall	F1-Score
Computer Science	Decision Tree	0.38 [0.32, 0.45]	0.47 [0.42, 0.53]
	Random Forest	0.40 [0.35, 0.45]	0.48 [0.43, 0.53]
	XGBoost	0.43 [0.38, 0.48]	0.51 [0.46, 0.55]
	Logistic Regression	0.41 [0.32, 0.49]	0.49 [0.41, 0.56]
Languages	Decision Tree	0.31 [0.25, 0.37]	0.40 [0.34, 0.45]
	Random Forest	0.36 [0.33, 0.40]	0.45 [0.42, 0.47]
	XGBoost	0.36 [0.32, 0.39]	0.44 [0.42, 0.47]
	Logistic Regression	0.36 [0.33, 0.39]	0.44 [0.42, 0.47]
Information Technology	Decision Tree	0.46 [0.35, 0.58]	0.50 [0.43, 0.58]
	Random Forest	0.58 [0.52, 0.64]	0.56 [0.53, 0.60]
	XGBoost	0.56 [0.51, 0.61]	0.56 [0.52, 0.59]
	Logistic Regression	0.65 [0.54, 0.76]	0.60 [0.55, 0.64]

The repeated splits substantially qualify the single-split results. Some values from the original partitions differ appreciably from the corresponding means across repetitions: for example, Decision Tree *dropout* recall increases from 0.30 to a mean of 0.46 for Information Technology and from 0.23 to 0.31 for Languages, whereas XGBoost decreases from 0.49 to 0.43 for Computer Science. More importantly, the confidence intervals overlap extensively within each program, indicating that the apparent model rankings from individual splits are not stable.

This instability is also reflected in how often each model obtains the highest F1-score. In Computer Science, XGBoost ranks first in 40% of the repeated splits, compared with 16% for Decision Tree and 4% for Random Forest; in Languages, the corresponding proportions are 24%, 10%, and 38%. Thus, the ordering observed in a single partition should not be interpreted as evidence that a particular tree-based algorithm is intrinsically better suited to a particular program.

The logistic regression baseline further supports this conclusion. Its confidence intervals overlap those of the tree-based models in Computer Science and Languages, while and

in Information Technology it obtains the highest mean *dropout* recall (0.65) and F1-score (0.60), achieving the highest F1-score in 64% of the repeated splits. These results provide little evidence that the additional complexity of the non-linear models yields a consistent advantage at the program level. Instead, the program-level comparisons emphasize the uncertainty associated with modeling relatively small institutional subsets.

Appendix B.3. Comparison Across Training Configurations

Table A5 consolidates the performance of the three tree-based models across the training configurations examined in the preceding subsections. The comparison highlights three recurring patterns: the effect of temporal evaluation, the relationship between class prevalence and class-specific recall, and the shift in predictive behavior produced by class rebalancing.

Table A5. Summary of performance on the *dropout* class across all evaluated training configurations. Precision, recall and F1-score refer to the *dropout* class, while accuracy is a model-level metric. Prevalence denotes the proportion of *dropout* students in the test partition.

Scenario	Prevalence	Model	Accuracy	Precision	Recall	F1-Score
Scenario 1: Complete dataset	0.36	Decision Tree	0.66	0.57	0.31	0.40
		Random Forest	0.67	0.60	0.32	0.42
		XGBoost	0.69	0.62	0.37	0.46
Scenario 2: Complete generations	0.36	Decision Tree	0.68	0.58	0.34	0.43
		Random Forest	0.69	0.62	0.35	0.45
		XGBoost	0.70	0.62	0.41	0.49
Scenario 3: Temporal split	0.43	Decision Tree	0.59	0.65	0.13	0.22
		Random Forest	0.58	0.69	0.07	0.13
		XGBoost	0.61	0.70	0.18	0.28
Scenario 4: Single center (CTC)	0.37	Decision Tree	0.64	0.53	0.30	0.38
		Random Forest	0.65	0.59	0.25	0.36
		XGBoost	0.66	0.57	0.32	0.41
Scenario 4: Single center (CES)	0.60	Decision Tree	0.62	0.64	0.84	0.72
		Random Forest	0.63	0.63	0.88	0.74
		XGBoost	0.63	0.65	0.83	0.73
Scenario 5: Computer Science	0.43	Decision Tree	0.63	0.62	0.40	0.49
		Random Forest	0.62	0.59	0.38	0.46
		XGBoost	0.65	0.62	0.49	0.55
Scenario 5: Languages	0.44	Decision Tree	0.59	0.60	0.23	0.34
		Random Forest	0.63	0.62	0.41	0.49
		XGBoost	0.63	0.61	0.40	0.48
Scenario 5: Information Technology	0.53	Decision Tree	0.51	0.56	0.30	0.39
		Random Forest	0.51	0.53	0.58	0.55
		XGBoost	0.53	0.55	0.54	0.55
Oversampling	0.36	Decision Tree	0.62	0.48	0.59	0.53
		Random Forest	0.64	0.50	0.59	0.54
		XGBoost	0.64	0.51	0.61	0.55
Undersampling	0.36	Decision Tree	0.62	0.48	0.60	0.54
		Random Forest	0.63	0.49	0.63	0.55
		XGBoost	0.64	0.51	0.64	0.57

Across the two institution-wide random-split scenarios, restricting the data to complete generations produces modest improvements for all three models. *Dropout* recall increases

from 0.31–0.37 on the complete dataset to 0.34–0.41 on the complete-generation subset, despite the reduction from 66,820 to 42,249 records. This pattern is consistent with the presence of outcome uncertainty in incomplete generations, although the magnitude of the improvement remains limited.

The temporal configuration produces the clearest deterioration in performance. Despite a higher *dropout* prevalence in the test partition (0.43), recall falls to 0.07–0.18 and F1-score to 0.13–0.28. Thus, the lower performance cannot be attributed to a lower representation of the positive class in the test data. Together with the rolling-origin analysis reported in Scenario 3, this result shows that performance estimated from random train-test partitions does not transfer directly to later admission cohorts.

Nevertheless, class prevalence is strongly associated with the operating behavior observed at the default threshold. In CTC, where *dropout* prevalence is 0.37, *dropout* recall ranges from 0.25 to 0.32; in CES, where prevalence reaches 0.60, recall rises to 0.83–0.88. A similar, although less pronounced, pattern appears in Information Technology, where the test-set prevalence is 0.53 and the ensemble models attain recalls of 0.54–0.58. These comparisons show that class-specific recall at the default threshold must be interpreted in conjunction with the outcome distribution of the population being evaluated.

Program-level results should be interpreted more cautiously. Although the single-split results in Table A5 suggest differences among algorithms and programs, the repeated-split analysis in Table A4 showed substantial overlap in confidence intervals and unstable model rankings. The program-level configurations therefore provide evidence of greater estimation uncertainty in these smaller subsets rather than a stable ordering of algorithms across individual programs.

Finally, rebalancing produces the largest increase in *dropout* recall among the non-temporal institution-wide configurations. Relative to Scenario 1, recall increases from 0.31–0.37 to 0.59–0.61 with oversampling and 0.60–0.64 with undersampling, while precision and accuracy decrease. This result demonstrates that the operating behavior of the classifiers can be shifted substantially toward identifying more students in the *dropout* class. However, recall at a fixed threshold does not establish whether the underlying discrimination has improved; this question is examined in Section 3.5 using threshold-independent discrimination and calibration measures.

Appendix C. Sensitivity to the Treatment of Undeclared Values

The temporal changes in the distribution and predictive importance of *not declared* values identified in Section 3.7.1 motivated a sensitivity analysis of their treatment. Table A6 reports *dropout* recall for the original treatment and the three alternative treatments defined in Section 2.9, for both the complete dataset and the temporal split.

On the complete dataset, retaining *not declared* as a valid category yields the highest *dropout* recall for all three models. Mode imputation produces only small reductions, whereas removing records containing undeclared values from the training partition reduces recall from 0.31 to 0.22 for Decision Tree, from 0.32 to 0.20 for Random Forest, and from 0.37 to 0.25 for XGBoost. Removing these records from both partitions produces similarly lower values. The corresponding PR-AUC lift also decreases by approximately 0.10–0.13 when the affected training records are discarded. These results provide empirical support for retaining *not declared* as a separate category rather than discarding a substantial portion of the available data.

The temporal split provides the more important robustness check. As shown in Section 3.7.1, the prevalence of *not declared* changes substantially between the older and more recent cohorts. Nevertheless, removing all affected records from the training partition reduces its size from 31,946 to 15,407 students without materially changing the low *dropout*

recall: the values change from 0.13 to 0.13 for Decision Tree, from 0.07 to 0.06 for Random Forest, and from 0.18 to 0.16 for XGBoost. PR-AUC lift changes by at most 0.05, and removing affected records from both the training and test partitions produces essentially the same pattern. Thus, the poor temporal performance persists even when records containing undeclared values are excluded, indicating that the temporal degradation cannot be attributed solely to the changing prevalence of the *not declared* category.

Table A6. *Dropout* recall under alternative treatments of undeclared values.

Scenario	Treatment	Train n	Decision Tree	Random Forest	XGBoost
Complete dataset	Retained as a category	53,456	0.31	0.32	0.37
	Imputed with the mode	53,456	0.29	0.32	0.35
	Dropped from training	39,491	0.22	0.20	0.25
	Dropped from both	39,491	0.23	0.24	0.28
Temporal split	Retained as a category	31,946	0.13	0.07	0.18
	Imputed with the mode	31,946	0.23	0.15	0.25
	Dropped from training	15,407	0.13	0.06	0.16
	Dropped from both	15,407	0.13	0.06	0.16

Mode imputation produces a different effect under the temporal split. *Dropout* recall increases from 0.13 to 0.23 for Decision Tree, from 0.07 to 0.15 for Random Forest, and from 0.18 to 0.25 for XGBoost. However, this improvement is not accompanied by better threshold-free discrimination: PR-AUC lift changes by only $+0.006$, -0.017, and -0.020, respectively. The higher recall therefore reflects a change in model behavior at the fixed decision threshold rather than an improvement in the models' ability to rank students by dropout risk. This result is consistent with the distinction established in Section 3.5 between discrimination and performance at a particular operating point.

The sensitivity analysis supports two conclusions. First, retaining undeclared values as a separate category is preferable to discarding the affected records in the complete-dataset setting, where their removal substantially reduces the available training data and predictive performance. Second, although the recording of *not declared* changes markedly across cohorts, alternative treatments of these values do not remove the temporal performance degradation. The temporal results should therefore be interpreted as a broader distribution-shift problem rather than as an artifact that can be explained by this single data-quality issue.

References

1. Kang, K.; Wang, S. Analyze and Predict Student Dropout From Online Programs. In *Proceedings of the International Conference on Computing and Data Analysis (ICCDA 2018), DeKalb, IL, USA, 23–25 March 2018*; ACM: New York, NY, USA, 2018. [CrossRef]
2. Araújo, E.J.M. Evasão no PROEJA: Estudo das Causas no Instituto Federal de Educação, Ciência e Tecnologia do Maranhão/IFMA. Master's Thesis, Universidade Católica de Brasília, Brasília, Brazil, 2012.
3. Tinto, V. Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Rev. Educ. Res.* **1975**, *45*, 89–125. [CrossRef]
4. Da Silva, F.C. Variáveis para Modelos Preditivos à Evasão na Educação Superior. Ph.D. Thesis, Universidade Federal de Santa Catarina, Florianópolis, Brazil, 2021.
5. da Cruz, R.C.; Juliano, R.C.; Souza, F.C.M.; Souza, A.C.C. A Score Approach to Identify the Risk of Student Dropout: An Experiment with Information Systems Course. In *Proceedings of the XIX Brazilian Symposium on Information Systems (SBSI '23), Maceió, Brazil, 29 May–1 June 2023*; ACM: New York, NY, USA, 2023. [CrossRef]
6. Bandura, A. *Social Foundations of Thought and Action: A Social Cognitive Theory*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1986.
7. Lobo, M. Panorama da evasão no ensino superior brasileiro: Aspectos gerais das causas e soluções. *Assoc. Bras. Mantenedoras Ensino Super. Cad.* **2012**, *25*, 14.

8. Silva, F.C.d.; Cabral, T.L.d.O.; Pacheco, A.S.V. Evasão em cursos de graduação: Uma análise a partir do censo da educação superior brasileira. In Proceedings of the XVI Coloquio Internacional de Gestion Universitaria (CIGU), Arequipa, Peru, 23–25 November 2016.
9. Karabacak, E.S.; Yaslan, Y. Comparison of Machine Learning Methods for Early Detection of Student Dropouts. In *Proceedings of the 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkey, 13–15 September 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 376–381.
10. Anh, B.N.; Minh, T.N.; Giang, N.H.; Son, N.T.; Hai, N.Q.; Chien, B.D. An University Student Dropout Detector Based on Academic Data. In *Proceedings of the 2023 IEEE Symposium on Industrial Electronics & Applications (ISIEA), Kuala Lumpur, Malaysia, 15–16 July 2023*; IEEE: Piscataway, NJ, USA, 2023. [CrossRef]
11. Lu, J.; Mou, J.; Li, P. Interpretive Analyses of Learner Dropout Prediction in Online STEM Courses. In Proceedings of the 2023 IEEE Frontiers in Education Conference (FIE), College Station, TX, USA, 18–21 October 2023.
12. Lu, Y.; Yeom, S.; Ryu, R.; Kim, S.H. Course-Level Clustering to Enhance Dropout Prediction Accuracy. *IEEE Access* **2025**, *13*, 204996–205013. [CrossRef]
13. Tete, M.F.; Sousa, M.d.M.; Santana, T.S.d.; Fellipe, S. Aplicação de métodos preditivos em evasão no ensino superior: Uma revisão sistemática da literatura. *Educ. Policy Anal. Arch.* **2022**, *30*, 149. [CrossRef]
14. Cho, C.H.; Yu, Y.W.; Kim, H.G. A Study on Dropout Prediction for University Students Using Machine Learning. *Appl. Sci.* **2023**, *13*, 12004. [CrossRef]
15. Villar, A.; de Andrade, C.R.V. Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discov. Artif. Intell.* **2024**, *4*, 2. [CrossRef]
16. Glandorf, D.; Lee, H.R.; Orona, G.A.; Pumptow, M.; Yu, R.; Fischer, C. Temporal and Between-Group Variability in College Dropout Prediction. In Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24), New York, NY, USA, 18–22 March 2024. [CrossRef]
17. Aakool, A.A.K.; Redha, D.A.; Atee, H.A. Mining Educational Data to Predict Influential Factors Patterns for Student Dropout in Iraq. *SSRG Int. J. Comput. Sci. Eng.* **2025**, *12*, 43–49. [CrossRef]
18. Kuntintara, W.; Warabuntaweesuk, P.; Rattapasakorn, S. Student Dropout Prediction Using Machine Learning. In Proceedings of the 2024 9th International Conference on Business and Industrial Research (ICBIR), Bangkok, Thailand, 23–24 May 2024.
19. Carballo-Mendívil, B.; Arellano-González, A.; Ríos-Vázquez, N.J.; Lizardi-Duarte, M.d.P. Predicting Student Dropout from Day One: XGBoost-Based Early Warning System Using Pre-Enrollment Data. *Appl. Sci.* **2025**, *15*, 9202. [CrossRef]
20. Realinho, V.; Machado, J.; Baptista, L.; Martins, M.V. Predicting Student Dropout and Academic Success. *Data* **2022**, *7*, 146. [CrossRef]
21. Barros, B.D.M. Aprendizado de Máquina Automático Aplicado à Predição da Evasão no Ensino Superior. Master's Thesis, Universidade Federal de Goiás (UFG), Goiânia, Brazil, 2022.
22. Fernández-García, A.J.; Preciado, J.C.; Melchor, F.; Rodriguez-Echeverria, R.; Conejero, J.M.; Sánchez-Figueroa, F. A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data. *IEEE Access* **2021**, *9*, 133076–133090. [CrossRef]
23. Bergsma, W. A bias-correction for Cramér's V and Tschuprow's T. *J. Korean Stat. Soc.* **2013**, *42*, 323–328. [CrossRef]
24. Nadeau, C.; Bengio, Y. Inference for the Generalization Error. *Mach. Learn.* **2003**, *52*, 239–281. [CrossRef]
25. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *Introduction to Statistical Learning: With Applications in R*; Springer: New York, NY, USA, 2021.
26. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, USA, 1984.
27. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
28. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; ACM: New York, NY, USA, 2016; pp. 785–794.
29. Jiménez, O.; Jesús, A.; Wong, L. Model for the Prediction of Dropout in Higher Education in Peru Applying Machine Learning Algorithms: Random Forest, Decision Tree, Neural Network, and Support Vector Machine. In Proceedings of the 2023 33rd Conference of Open Innovations Association (FRUCT), Žilina, Slovakia, 24–26 May 2023.
30. Costa, A.G.; Mattos, J.C.B.; Primo, T.T.; Cechinel, C.; Muñoz, R. Model for Prediction of Student Dropout in a Computer Science Course. In Proceedings of the 2021 XVI Latin American Conference on Learning Technologies (LACLO), Virtually, 19–21 October 2021. [CrossRef]
31. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 559–563.
32. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]

33. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.
34. Hardt, M.; Price, E.; Srebro, N. Equality of Opportunity in Supervised Learning. In Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016), Barcelona, Spain, 5–10 December 2016; pp. 3315–3323.
35. Chouldechova, A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef] [PubMed]
36. Kleinberg, J.; Mullainathan, S.; Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Berkeley, CA, USA, 9–11 January 2017; pp. 43:1–43:23.
37. Nagy, M.; Molontay, R. Predicting Dropout in Higher Education Based on Secondary School Performance. In Proceedings of the 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), Las Palmas de Gran Canaria, Spain, 21–23 June 2018; pp. 389–394.
38. Berens, J.; Schneider, K.; Görtz, S.; Oster, S.; Burghoff, J. Early Detection of Students at Risk–Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *J. Educ. Data Min.* **2019**, *11*, 1–41. [CrossRef]
39. Mduma, N. Data Balancing Techniques for Predicting Student Dropout Using Machine Learning. *Data* **2023**, *8*, 49. [CrossRef]