

Student Transfer Prediction under Class Imbalance: Evaluation, Class Weighting, and Feature Importance

Branson Witt

BWitt@clarku.edu

Clark University

Worcester, Massachusetts, USA

Li Han*

LHan@clarku.edu

Clark University

Worcester, Massachusetts, USA

Abstract

Student transfer is a consequential form of attrition in higher education, affecting students, institutions, and local communities. Prior work has shown that machine learning models can predict student retention and transfer using institutional, behavioral, and social engagement data. However, transfer prediction is an imbalanced classification problem: transfer students represent the minority class, yet they are the class of greatest practical interest for early support. In this study, we revisit student transfer prediction using a large multi-institutional dataset, focusing on evaluation metrics, class-weighted modeling, and feature-importance analysis under class imbalance. We show that accuracy alone can obscure weak minority-class detection, and adopt recall and precision-recall area under the curve as intervention-oriented evaluation metrics. We further show that increasing the transfer-class weight improves recall and reduces false negatives, at the cost of additional false positives, providing institutions with a practical mechanism for tuning models according to the relative costs of missed support opportunities versus unnecessary outreach. Finally, we compare feature-importance metrics using top-feature similarity and overall correlation, showing that Spearman correlation is an effective pre-training filter and that key behavioral and academic features are consistently identified across metrics and models.

CCS Concepts

• **Applied computing** → **Education**; • **Computing methodologies** → *Feature selection*; Supervised learning by classification; Boosting; • **Information systems** → Data cleaning.

Keywords

Student transfer, student retention, machine learning, evaluation metrics, class imbalance, feature importance analysis.

ACM Reference Format:

Branson Witt and Li Han. 2026. Student Transfer Prediction under Class Imbalance: Evaluation, Class Weighting, and Feature Importance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and*

*Faculty advisor from Department of Computer Science, Clark University

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '26, Jeju, Korea

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0400-5/25/08

<https://doi.org/XXXXXXX.XXXXXXX>

Data Mining (KDD '26). ACM, New York, NY, USA, 7 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Student retention in postsecondary education remains a pressing problem worldwide. Higher education institutions are increasingly competing for smaller cohorts of students as enrollment at degree-granting postsecondary institutions decreased by 15% from 2010 to 2021 [1]. This is occurring at the same time as high post-pandemic inflation and economic uncertainty drive up operating costs for many postsecondary education institutions, leading to program cuts and even permanent closures [2]. This contraction in student supply makes retention increasingly critical for institutional sustainability.

In total, 41% of people graduate on time and 60% of people graduate within 6 years at a typical 4-year degree program within the United States [3]. For a typical 2-year degree program, the graduation rates within 3 years from the program are around 30% [3]. By not graduating, students incur both sunk costs, such as student loans and tuition, and opportunity costs, including lost future wages with a degree. Those with a bachelor's degree earn approximately \$800,000 more in lifetime earnings than those without [4], underscoring the high opportunity cost of dropping out.

One major and often overlooked form of student attrition is transfer. Studies show that 37.1% of students in a 2-year degree program and 38.5% of students in a 4-year degree program will transfer at some point while working toward their degree, with 94.4% of two-year transfers doing so before earning any credential [5]. On average, a student who transfers will lose 1 semester's worth of credits and incur an additional \$13,081 to \$26,396 in attendance costs [4]. For postsecondary institutions, attrition, such as transfers, can still entail sunk costs and lost revenue [6]. These facts underscore the importance of postsecondary degree institutions having early, reliable methods to identify at-risk students so they can intervene with targeted programs and resources. This is especially true, as approximately a quarter of all students leave within their first year [7], and the number of transfers is increasing [8]. The volume of research on early student performance prediction has likewise grown in recent years, reflecting the widespread interest in data-driven approaches to retention [9].

Prior research on student retention has long emphasized that persistence is shaped by both individual academic preparation and students' integration into the institutional community. Foundational theories of student departure highlight the importance of academic and social integration in students' decisions to remain enrolled [10–13], while empirical studies have linked retention to prior academic achievement, demographic and socioeconomic factors, peer relationships, sense of belonging, and students' positions

in social networks [14; 15]. More recently, educational data mining and learning analytics have shown that machine learning models can support early identification of at-risk students using administrative, academic, behavioral, and LMS interaction data [16–18].

The two most relevant prior studies provide the foundation for our work. Matz et al. [19] introduced a large multi-institutional dataset that combines socio-demographic, academic, and app-based engagement features from more than 50,000 students and demonstrated that machine learning models can predict student retention with strong performance when institutional and behavioral engagement data are combined. Building on this dataset, Goka et al. [20] further examined student transfer prediction across multiple universities, evaluated a broader set of machine learning models, and identified important institutional, application engagement, community engagement, and social network features associated with transfer behavior. Together, these studies show that student transfer can be predicted from multidimensional student data and that behavioral and social engagement features provide meaningful information beyond traditional institutional records.

Our work extends this line of research by focusing on evaluation and model design choices that are especially important when student transfer prediction is used to support institutional intervention in the presence of class imbalance. Instead of proposing a new predictive model, this research presents a tunable framework for deploying and evaluating transfer prediction models to help universities identify at-risk students and intervene. First, we examine evaluation metrics for transfer prediction, emphasizing recall, PR AUC (Precision-Recall Area Under the Curve), false negatives, and false positives as complements to commonly reported metrics such as accuracy and ROC AUC (Receiver Operating Characteristic - Area Under the Curve). This framing reflects the intervention setting: false negatives correspond to missed opportunities to provide timely support to students who may transfer, while false positives correspond to advising or outreach resources directed toward students who would not ultimately transfer. Second, we study class-weighted modeling as a practical mechanism for tuning this tradeoff according to institutional priorities. Rather than treating the class-ratio weight as automatically optimal, we evaluate how different transfer-class weights affect recall and the balance between false negatives and false positives. Third, we conduct feature-importance analysis by comparing top-feature similarity and overall correlation across importance metrics, and by testing whether compact feature subsets selected by these metrics can preserve or improve recall. Finally, we show that Spearman’s correlation is a computationally inexpensive method for filtering unpromising features before training a model. Together, these analyses shift the focus from overall predictive accuracy toward class-imbalance-aware, tunable models aligned with institutional intervention priorities.

2 Dataset, Preprocessing, and Model Selection

The dataset used in this paper was introduced by Matz et al. [19] and comprises data from 50,095 students collected via the READY application across 4 universities. The data cover a broad range of institutional characteristics, including demographics, financial status, and student achievement, as well as descriptive statistics of app engagement and usage, including mean, standard deviation,

and skewness. The target variable is retention status, taking the values Continued or Transferred. The dataset is publicly available at <https://osf.io/bhaqp> [21].

Universities 1 through 3 are traditional university campuses; University 4 is a single entity comprising 16 community colleges. All universities have 462 features split into institution-reported and app-derived behavioral features. We drop University 1 for having a smaller comparable sample size ($n = 476$) and drop University 2 due to a lack of behavioral variation across models with different parameters. University 2 proved to produce the same top features and scores among metrics across a wide range of boosted tree models. The remaining Universities 3 and 4 provide a solid basis for testing model parameters. University 3 comprises 2,547 students with an average GPA of 3.02 and a transfer rate of 15.8%; University 4 comprises 45,062 students with an average GPA of 2.74 and a transfer rate of 10.2%.

For preprocessing, features with greater than 70% missing values were removed, as were constant features with a single unique entry. After preprocessing, 244 features remain for University 3 and 233 for University 4.

Prior work on this dataset by Goka et al. [20] found that logistic regression and boosted tree models achieve the best performance. Accordingly, we explored logistic regression, XGBoost [22], LightGBM [23], and CatBoost [24]. Logistic regression was subsequently removed due to poor performance, and LightGBM was removed as it exhibited behavior similar to XGBoost.

CatBoost was configured using two separate Pool objects to exploit its internal handling of categorical data. XGBoost was trained on a one-hot encoded dataframe. Both models were trained using LogLoss with a 5-fold stratified cross-validation split, retaining the fold with the highest average precision score as the final model. CatBoost was trained for up to 500 iterations with a learning rate of 0.05, a maximum tree depth of 8, and early stopping after 50 iterations. XGBoost used 300 estimators with the same learning rate and tree depth, but without early stopping. Baseline performance, evaluated on a 10% holdout, is reported in Table 1. As discussed in the following section, these results reveal a pattern common to imbalanced classification: high overall accuracy alongside substantially weaker minority-class detection.

3 Evaluation under Class Imbalance

Student transfer prediction is an imbalanced binary classification problem. In the two universities analyzed in this study, transfer students represent the minority class: 15.8% in University 3 and 10.2% in University 4. As a result, accuracy alone can be misleading: a model may obtain high accuracy by primarily predicting the majority class, continued enrollment, while still missing many students who eventually transfer.

This pattern is visible in Table 1. Despite accuracy scores of 0.85 and 0.93 for Universities 3 and 4 respectively, CatBoost and XGBoost achieve recall values below 0.40 across all configurations, meaning more than 60% of students who ultimately transfer are missed entirely. The gap is especially stark for CatBoost on University 3, where a recall of 0.16 means the model correctly identifies fewer than one in six transfer students. High specificity values (above 0.96 in all cases) confirm that both models are effectively optimized

Table 1: Performance of Unweighted Models

Metric	CatBoost		XGBoost	
	Univ. 3	Univ. 4	Univ. 3	Univ. 4
Accuracy	0.8471	0.9268	0.8471	0.9277
Precision	0.7000	0.8045	0.6000	0.8382
Recall	0.1628	0.3815	0.2791	0.3685
Specificity	0.9858	0.9894	0.9623	0.9918
F1	0.2642	0.5175	0.3810	0.5120
ROC AUC	0.7922	0.7912	0.7952	0.7762
PR AUC	0.5036	0.5314	0.5030	0.5239

for the majority class. These results illustrate why accuracy and specificity are insufficient as primary evaluation metrics in this setting, and motivate the use of recall-based measures to more directly align with the goal of identifying at-risk students before they leave.

Prior work using this dataset has reported strong performance using ROC AUC and accuracy as primary metrics [19; 20], which is well-suited to evaluating overall classification performance. The ROC curve plots the true positive rate against the false positive rate across classification thresholds, evaluating performance across both classes. The precision-recall curve, by contrast, plots positive predictive value against true positive rate and focuses on the positive class, making it more sensitive to how well a model identifies the minority class [25–27]. PR AUC summarizes this curve as a single score: a higher value indicates that the model maintains both high precision and high recall across thresholds. PR AUC was estimated using the average precision score during our model evaluation. This issue is not fully settled, however: Richardson et al. argue that ROC AUC is robust to class imbalance, whereas PR AUC is itself sensitive to positive-class prevalence [28].

In the intervention setting that motivates this work, we adopt recall and PR AUC as our primary evaluation metrics. False negatives correspond to missed opportunities to support students who may transfer, while false positives correspond to outreach directed at students who would not. Recall and PR AUC directly reflect performance on this minority class and are therefore most aligned with the practical goal of early identification.

4 Class-Weighted Modeling

Class-weighted training penalizes the misclassification of the minority class more heavily during optimization, providing a direct and interpretable mechanism for improving recall without modifying the data distribution. Critically, the weight assigned to the positive class need not be fixed at the class-ratio value; it can be tuned to reflect institutional priorities regarding the relative costs of false negatives and false positives.

To address class imbalance, we increased the weight of the positive class, which represents transfer students. Figure 1 shows how varying the positive class weight changes the number of false negatives and false positives for CatBoost and XGBoost in Universities 3 and 4. The conventional class-balancing weights, based on the ratio of non-transfer to transfer students, are 5.32 for University 3 and 8.81 for University 4. However, our results indicate that these values

need not be treated as optimal. Increasing the positive class weight beyond the class-ratio value can further reduce false negatives and improve recall, although at the cost of more false positives.

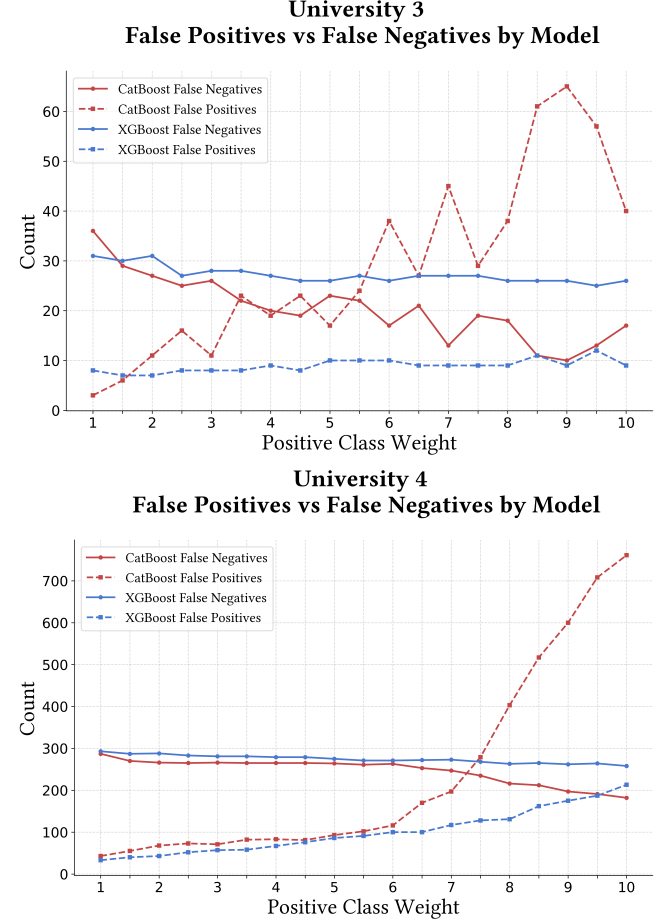


Figure 1: False negative and false positive counts by model across varying positive class weights.

Table 2 confirms this tradeoff. Increasing the positive class weight consistently improves recall, especially for CatBoost. For University 3, CatBoost recall increases from 0.3721 at weight 2.00 to 0.6047 at weight 10.00; for University 4, it increases from 0.4267 to 0.6078. XGBoost follows the same trend, though with smaller changes. This recall improvement comes at a cost: PR AUC tends to decrease as the positive class weight increases, reflecting the additional false positives introduced as more students are classified as likely transfers, though this trend is not strictly monotonic across all configurations. These results suggest that class weighting is an effective and interpretable approach to tune the aggressiveness with which a model identifies transfer-risk students. The appropriate weight should therefore depend on institutional priorities: higher weights may be appropriate when missed transfer students are especially costly, while lower weights may be preferable when interventions are expensive or potentially stigmatizing.

Table 2: Performance of Weighted Models

University	Pos. Weight	Model	Recall	PR AUC
Univ. 3	2.00	CatBoost	0.3721	0.5425
		XGBoost	0.2791	0.4898
	5.32	CatBoost	0.5116	0.4976
		XGBoost	0.3488	0.5050
	10.0	CatBoost	0.6047	0.4960
		XGBoost	0.3953	0.5160
Univ. 4	2.00	CatBoost	0.4267	0.5286
		XGBoost	0.3793	0.5239
	8.81	CatBoost	0.5819	0.5298
		XGBoost	0.4310	0.5055
	10.0	CatBoost	0.6078	0.5247
		XGBoost	0.4440	0.5048

Figure 1 further shows that CatBoost and XGBoost respond differently to class weighting. In both universities, CatBoost is more sensitive to changes in the positive class weight than XGBoost, producing larger changes in false-negative and false-positive counts as the weight increases. XGBoost changes more gradually, with false positives increasing and false negatives decreasing at a slower rate. This pattern is especially clear for University 4, where CatBoost more consistently reduces false negatives while increasing false positives. For University 3, CatBoost shows less monotonic behavior, although the overall trend is similar. This instability may be partly due to the much smaller holdout set for University 3 ($N = 255$) compared with University 4 ($N = 4,507$).

5 Feature Importance Analysis

Feature-importance analysis serves two related purposes in this study. First, it helps assess whether different importance metrics point to similar predictors of student transfer, which is important because individual importance measures can reflect the assumptions of a particular model or scoring method. Second, it provides a basis for evaluating whether smaller feature subsets can retain useful predictive performance, especially recall for the minority transfer class. We therefore examine both the consistency of feature-importance results across metrics and the predictive usefulness of selected features.

5.1 Top-Feature Similarity and Overall Correlation Across Importance Metrics

We compare feature-importance metrics with two complementary measures: Jaccard similarity among the top-ranked features of each metric and Spearman correlation across the full feature orderings produced by each metric. The Jaccard analysis assesses whether different metrics identify the same important features, whereas the Spearman analysis captures broader similarity in how metrics order the full feature set. More specifically, Jaccard similarity between two sets is defined as the size of their intersection divided by the size of their union, ranging from 0 (no overlap) to 1 (identical sets) [29]. Spearman rank correlation measures the monotonic relationship

between two rankings, ranging from -1 (perfectly inverse) to $+1$ (perfectly concordant), with 0 indicating no monotonic association.

We consider three categories of importance metrics. The first comprises traditional correlation measures: Pearson’s ρ , which applies to interval variables only; Spearman’s ρ , which applies to interval and ordinal variables; and PhiK [30], which applies to interval, ordinal, and categorical variables. Because the dataset contains categorical features, not all variables can be ranked using Pearson’s or Spearman’s ρ .

The second category comprises model-based importance metrics. For XGBoost, we use gain, defined as the average improvement in model performance from using a feature in a split, and the weight, the number of times a feature is used in a decision node across all trees [31]. For CatBoost, we use the prediction value change, the average change in predicted probability when a feature value changes, and loss function change, the difference in the model’s loss with and without the feature [32]. The third category comprises SHapley Additive exPlanations (SHAP) values, which quantify individual feature contributions to model predictions [33], computed separately for both XGBoost and CatBoost.

To assess agreement between metrics, Jaccard’s similarity index was calculated pairwise between the top 5 features for each metric (Figure 2), and Spearman’s rank correlation was calculated pairwise between full feature rankings (Figure 3).

When analyzing agreement among feature importance metrics, patterns emerge across metric types and models. Across most universities and agreement metrics, Spearman’s ρ and Pearson’s ρ are highly consistent. Interestingly, this pattern is absent for Jaccard’s index at University 3. Model-specific metrics also tend to agree with one another. For example, CatBoost’s SHAP, prediction value change, and loss function change are all in close agreement between measurements and universities. The same pattern holds for XGBoost SHAP and, to a lesser extent, for XGBoost weights, but is not observed for XGBoost gain. CatBoost-based metrics also agree well with XGBoost SHAP value importance, but the reverse is not true. For University 4, Jaccard’s index demonstrates agreement among all top correlation-based importance metrics. The most important finding is that Spearman’s ρ is a moderate predictor of the top CatBoost features, as indicated by Jaccard’s index for both universities. This can be used to identify a subset of predictors for CatBoost models before training, thereby filtering out poor-performing features.

5.2 Prediction with Selected Features

To further evaluate the feature-importance metrics, we trained XGBoost and CatBoost models on the top 25, 50, and 100 features selected by each metric and assessed recall at both universities. Model performance was also evaluated using a majority vote among the nine importance metrics, where a feature was included if selected by the majority. Because the majority vote aggregates across all metrics, the resulting feature pool may be larger than those from any individual metric at higher selection thresholds. All models were trained with a positive class weight of 2. Results are reported in Table 3, with PhiK excluded due to poor performance. Bold values indicate improved recall over the weighted baseline in Table 2.

Table 3 shows that feature-subset models do not consistently outperform the weighted baseline, but targeted selection can yield

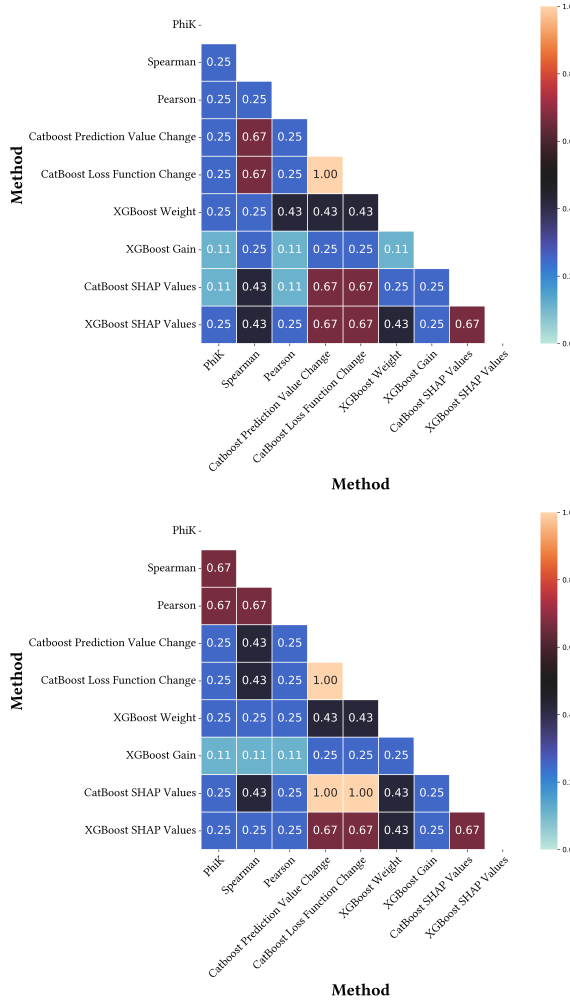


Figure 2: Pairwise Jaccard similarity index between importance metrics based on their top 5 most important features for University 3 (top) and University 4 (bottom) (Visualization approach partially adapted from [34])

meaningful gains. The highest recall value in the table is achieved with the top 50 features from Spearman correlation in a CatBoost model on University 3, reaching 0.4651, a 25% increase over the weighted baseline. This result is consistent with the Jaccard similarity findings in Figure 2, which showed moderate agreement between Spearman correlation and CatBoost-based importance metrics. Together, these results support Spearman correlation as a computationally efficient pre-training filter for CatBoost models. However, this strategy may be less effective for datasets with a larger proportion of categorical variables, for which Spearman correlation cannot be estimated.

XGBoost improved recall more often when trained on features selected across all metrics, but its recall scores were generally lower than CatBoost's. Training either model on features identified by SHAP values led to similar performance regardless of which model

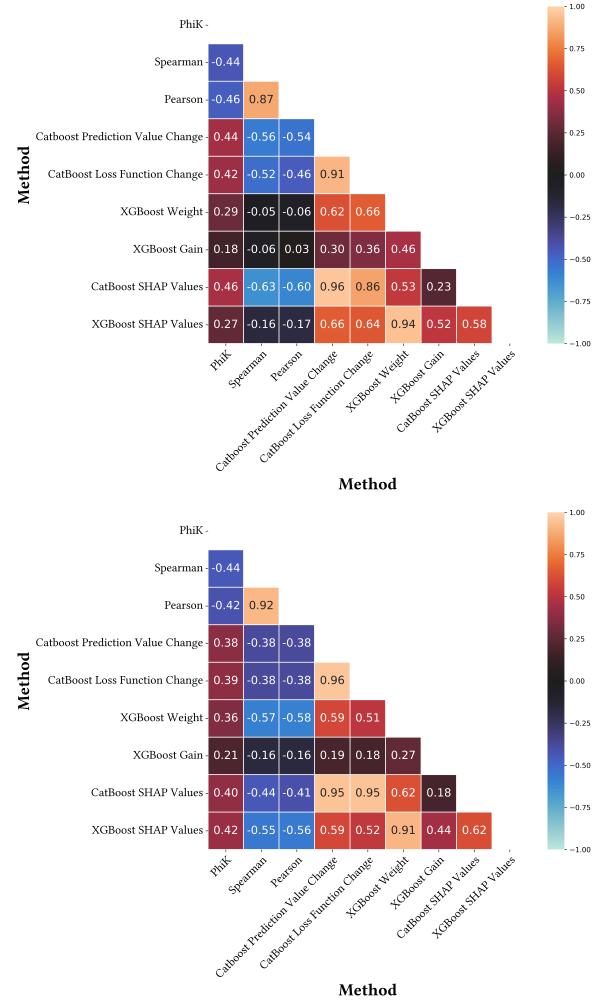


Figure 3: Spearman rank correlation between importance metric feature rankings for University 3 (top) and University 4 (bottom) (Visualization approach partially adapted from [34])

generated the SHAP rankings. Majority voting achieved competitive recall but typically did not surpass the top individual metrics, except for XGBoost on University 3 with the top 25 features. Pearson-selected features improved recall on University 3 for both models, but not on University 4, likely because University 4 relies more heavily on ordinal and categorical features that Pearson cannot rank.

The only feature ranking in the top 25 across all importance metrics and universities is cumulative GPA, which is by far the strongest predictor of transfer: lower GPAs are consistently associated with higher transfer likelihood, a pattern confirmed by SHAP values across all models and universities. Features ranking in the top 50 across all metrics and universities are:

- High School GPA

Table 3: Recall by Feature-Importance Metric and Subset Size

Metric	Top	University 3		University 4	
		CatBoost	XGBoost	CatBoost	XGBoost
CB Prediction Value Change	25	0.4186	0.3721	0.4289	0.4095
	50	0.4186	0.3256	0.4267	0.3966
	100	0.3953	0.2791	0.4246	0.4030
CB Loss Func. Change	25	0.4186	0.3953	0.4289	0.4030
	50	0.3721	0.2791	0.4224	0.4030
	100	0.4186	0.3256	0.4224	0.4030
XGBoost Weight	25	0.4186	0.3256	0.4246	0.4073
	50	0.3953	0.3256	0.4267	0.3987
	100	0.3721	0.3023	0.4289	0.4009
XGBoost Gain	25	0.3721	0.3721	0.4289	0.4138
	50	0.3721	0.2326	0.4289	0.3987
	100	0.3721	0.3488	0.4267	0.3966
CatBoost SHAP	25	0.4419	0.3953	0.4289	0.4095
	50	0.3721	0.3023	0.4289	0.4030
	100	0.3721	0.2791	0.4267	0.4030
XGBoost SHAP	25	0.3953	0.3256	0.4289	0.4052
	50	0.4419	0.3256	0.4267	0.4009
	100	0.3023	0.3023	0.4267	0.3987
Pearson	25	0.4419	0.3023	0.3707	0.3384
	50	0.4419	0.3023	0.3685	0.3319
	100	0.3721	0.3488	0.3642	0.3168
Spearman	25	0.4186	0.3023	0.4246	0.4159
	50	0.4651	0.2791	0.4289	0.3987
	100	0.3488	0.3488	0.4267	0.3987
Majority Vote	25	0.4186	0.4419	0.4267	0.4052
	50	0.3256	0.3256	0.4289	0.4009
	100	0.3488	0.3256	0.4289	0.4009

- Total number of clicks on the READY education app homepage
- Maximum time between two app sessions
- Average app session duration
- Total number of active days on the app
- Total number of active weekdays on the app
- Maximum time between two app actions on weekdays
- IQR of the time between two app actions
- Entropy of different senders in the app chat

Previous studies using the same dataset have identified cumulative GPA, high school GPA, and total homepage clicks as strong predictors of student transfer [19; 20]. Matz et al. [19] further found that total active days and total active weekdays are highly correlated, and Goka et al. identified sender entropy as an important predictor. While the remaining top features identified here differ from those in prior work, they often represent alternative statistical aggregations of the same underlying signals. For example, whereas a previous study used the standard deviation of time between sessions, we find that the maximum time between sessions is more predictive. This convergence on the same underlying behavioral signals, even across different aggregations, suggests that these features capture genuinely robust patterns in student behavior.

6 Model Framework in Action

For a university deploying a model for student transfer prediction in action, there are a few considerations and implementations to

make. First, model performance and validation should be done with recall and PR AUC as primary metrics, with accuracy serving as a validation metric to ensure models are not overly focused on the positive class. This allows for optimal detection of minority-class performance, as discussed throughout the paper. Secondly, a positive class weight should be applied when transfer students are a minority, with the positive class weight serving as an adjustable point rather than a fixed value that depends on shifting university priorities. The aim should be to increase true positives while managing the number of false positives, depending on the university's changing cost of missing a student transfer. Finally, universities can use Spearman's correlation as a pre-training filter when training a CatBoost model on their internal administrative and app data to reduce feature dimensionality before training. This is particularly useful for institutions with limited computing resources or a large number of candidate features from their student information systems.

7 Conclusion and Future Work

This paper examined evaluation metrics, class-weighted modeling, and feature-importance analysis for student transfer prediction in machine learning models. On evaluation, we showed that accuracy and specificity alone are insufficient for transfer prediction, as models achieve high overall accuracy while missing students who ultimately transfer since recall and PR AUC more directly reflect performance on the minority transfer class. Increasing the positive class weight improved recall and correctly identified more transfer students, at the cost of more false positives. This tradeoff means that universities should select personalized class-weight ratios based on institutional priorities and economic capabilities. Spearman correlation effective pre-training filter for CatBoost models, with cumulative GPA and other resulting top features being consistent with previous work.

This work opens several directions for future research. Further investigation into SHAP-based feature contributions and model explainability would help clarify how top features interact within the models. The choice of evaluation metrics under class imbalance also warrants continued study, as the relative merits of ROC AUC and PR AUC remain context-dependent and unsettled across varying class ratios and intervention settings. Identifying thresholds for key features such as cumulative GPA across a broader pool of universities would benefit institutions that lack the resources to train personalized models. As a longer-term direction, quantifying the economic value of a transferring student would better inform decisions about the acceptable level of resources to devote to false positives, requiring additional institutional data and economic modeling beyond the scope of this work.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Number 2436765. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

- [1] National Center for Education Statistics. NCES undergraduate enrollment fast facts. <https://nces.ed.gov/fastfacts/display.asp?id=98>, 2023.
- [2] Robert Kelchen, Dubravka Ritter, and Douglas Webber. Predicting college closures and financial distress. Working Paper WP 24-20, Federal Reserve Bank of Philadelphia, December 2024.
- [3] National Center for Education Statistics. Status and trends in the education of racial and ethnic groups: Indicator 23: Postsecondary graduation rates. https://nces.ed.gov/programs/raceindicators/indicator_red.asp, 2019.
- [4] William Carroll and Brenae Smith. The costs of today's college credit transfer system for learners and the mindsets and practices that reduce them. White paper, The Center for Higher Education Policy and Practice, May 2024.
- [5] Doug Shapiro, Afet Dundar, Faye Huie, Phoebe Khasiala Wakhungu, Ayesha Bhimdiwali, Angel Nathan, and Hye Youngsik. Transfer and mobility: A national view of student movement in postsecondary institutions, fall 2011 cohort. Signature Report 15, National Student Clearinghouse Research Center, Herndon, VA, July 2018.
- [6] Jane Wellman, Nate Johnson, and Patricia Steele. Measuring (and managing) the invisible costs of postsecondary attrition. Technical report, American Institutes for Research, September 2012. Commissioned by the Delta Cost Project.
- [7] U.S. Department of Education, National Center for Education Statistics. Integrated Postsecondary Education Data System (IPEDS). Table 326.30: Retention of first-time degree-seeking undergraduates at degree-granting postsecondary institutions, by attendance status, level and control of institution, and percentage of applications accepted: Selected years, 2012 through 2023. https://nces.ed.gov/programs/digest/d24/tables/dt24_326.30.asp, 2024.
- [8] S. Karamarkovich, B. Berg, J. Cohen, M. Holsapple, and D. Shapiro. Transfer enrollment and pathways: Fall 2024. Technical report, National Student Clearinghouse Research Center, Herndon, VA, March 2025.
- [9] Javier López-Zambrano, Juan Alfonso Lara Torralbo, and Cristóbal Romero. Early prediction of student learning performance through data mining: A systematic review. *Psicothema*, 33(3):456–465, 2021.
- [10] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125, 1975.
- [11] Vincent Tinto. *Completing College: Rethinking Institutional Action*. University of Chicago Press, 2012.
- [12] Alexander W. Astin. Student involvement: A developmental theory for higher education. *Journal of College Student Personnel*, 25(4):297–308, 1984.
- [13] John P. Bean. Nine themes of college student retention. In *College Student Retention: Formula for Student Success*, pages 215–243. 2005.
- [14] Scott L. Thomas. Ties that bind: A social network approach to understanding student integration and persistence. *The Journal of Higher Education*, 71(5):591–615, 2000.
- [15] Maithreyi Gopalan and Shannon T. Brady. College students' sense of belonging: A national perspective. *Educational Researcher*, 49(2):134–137, 2020.
- [16] Lovenoor Aulck, Dev Nambi, Nishant Velagapudi, Joshua Blumenstock, and Jevin West. Mining university registrar records to predict first-year undergraduate attrition. In *Proceedings of the 12th International Conference on Educational Data Mining*, 2019.
- [17] Or Goren, Itay Gelbart, and Avner Eizenberg. Early prediction of student dropout in higher education using machine learning models. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 349–359, 2024.
- [18] Anaile M. Rabelo, Paulo G. Costa, and Daniela A. dos Santos. A model for predicting dropout of higher education students. *Data Science and Management*, 8:72–85, 2025.
- [19] Sandra C. Matz, Christian S. Bukow, Helena Peters, Carol Deacons, Andrei Dinu, and Clemens Stachl. Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. *Scientific Reports*, 13, 2023.
- [20] Keerthana Goka, Kunal Malhan, Vinod Nithin Kumar Rachakonda, and Li Han. Leveraging machine learning to understand and predict student transfer in higher education. Unpublished manuscript; originally prepared for the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25), 2025.
- [21] Sandra C. Matz et al. Student retention dataset. <https://osf.io/bhaqp/>, 2023. Accessed May 2025.
- [22] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, USA, 2016.
- [23] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.
- [24] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, 2018.
- [25] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, 2006.
- [26] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015.
- [27] Faezeh Movahedi, Rema Padman, and James Antaki. Limitations of receiver operating characteristic curve on imbalanced data: Assist device mortality risk scores. *The Journal of Thoracic and Cardiovascular Surgery*, 165:1433–1442, 2021.
- [28] Eve Richardson, Raphael Trevizani, Jason A. Greenbaum, Hannah Carter, Morten Nielsen, and Bjoern Peters. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, 5(6):100994, 2024.
- [29] Oona Rainio, Jarmo Teuvo, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14:6086, 2024.
- [30] M. Baak, R. Koopman, H. Snoek, and S. Klous. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics & Data Analysis*, 152:107043, 2020.
- [31] XGBoost Developers. XGBoost documentation. <https://xgboost.readthedocs.io>.
- [32] CatBoost Developers. CatBoost documentation: Feature importance. <https://catboost.ai/docs/en/concepts/fstr#regular-feature-importance>.
- [33] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [34] Alan Inglis. Correlation heatmaps. <https://ajh1143.github.io/Corr/>. Accessed: 2025-04-28.

Received April 30, 2026