Identifying Interpretable Features Impacting Nontraditional Undergraduate Computer Science Student Retention

1st Jiang Li*
Computing Sciences and Mathematics
Franklin University
Columbus, Ohio U.S.A.
jiang.li2@franklin.edu

2nd Chunbo Chu*
Computing Sciences and Mathematics
Franklin University
Columbus, Ohio U.S.A.
chunbo.chu@franklin.edu

3rd Todd Whittaker
Computing Sciences and Mathematics
Franklin University
Columbus, Ohio U.S.A.
todd.whittaker@franklin.edu

***Abstract*—This research full paper describes an investigation of nontraditional undergraduate Computer Science (CS) students' unique retention issues by using machine learning. It presents new insight into the most critical factors impacting their retention compared to traditional students. CS student retention is a pressing concern for many universities. Traditional students have been well-studied, but the growing number of nontraditional students have received less attention. These students often face unique challenges due to their different backgrounds and experiences compared to their traditional counterparts. For example, they typically delay postsecondary enrollment, have dependents, and attend school part time. Existing research has limited information on how these varying characteristics influence their academic success. Consequently, applying retention strategies based solely on traditional student data might not be effective for nontraditional students. This study aims to bridge the knowledge gap regarding nontraditional undergraduate CS student retention. We proposed two key questions: (1) What are the primary factors influencing nontraditional undergraduate CS student attrition? and (2) How can we predicate students at a higher risk of dropping out at an earlier stage? We analyzed the data of nontraditional undergraduate students enrolled in the CS major at a university with a predominantly nontraditional student population. The data span 13 years and encompass demographic, academic, and behavioral features. We trained five machine learning models (Logistic Regression, Support Vector Machines, K-Nearest Neighbors, Random Forest, and Gradient Boosting Trees) and an ensemble model. By considering both accuracy and AUC (Area Under the Curve) of the optimized models, we found that overall, the Gradient Boosting Trees model performs the best. Using permutation importance in the Gradient Boosting Trees model, we found nine most significant features impacting student retention: (1) GPA. Students with higher GPA (cumulative or in the last semester) are more likely to retain. (2) Course load. Students taking more classes overall or in a given semester tend to retain better. (3) Number of face-to-face classes taken. Students taking fewer face-to-face classes and more online classes are more likely to retain. (4) Age. Younger students tend to retain better. (5) Whether students transferred through a community college partnership. Students recruited through a partnership with two-year institutions tend to retain better. (6) Financial aid. Students with financial aid are less likely to drop out than students without any financial aid. These innovative findings deepen the knowledge of the specific retention challenges faced by nontraditional undergraduate CS students.**

This study successfully addressed the proposed research questions, providing valuable insights into nontraditional undergraduate CS student retention. Identifying key influential features improves the interpretability of the predictive models, leading to better compliance with regulations and reduced concerns about bias. Additionally, the optimized model can inform interventions and student services to enhance retention rates.

***Index Terms*—nontraditional students, retention, machine learning, key features, model interpretability.**

I. Introduction

Retention is a major concern for many higher education institutes around the world, especially in science, technology, engineering, and mathematics (STEM) majors. In the United States, there are persistent challenges in producing and retaining STEM talent to meet the demands as the workforce shifts to more technology-based jobs. According to the National Center for Education Statistics, 48% of bachelor's degree students who entered STEM fields between 2003 and 2009 had left these fields by Spring 2009. Roughly half of these students switched to a non-STEM major, and the rest of them left college without earning any credentials [1].

Traditionally, a college student is envisioned as between the ages of 18 and 24, matriculating immediately after high school, living on campus, pursuing a full-time course load, attending daytime in-person classes, and having no dependents. However, this stereotype no longer accurately represents the diverse student body of modern days [2]. More students challenge this norm by exhibiting one or more alternative characteristics: they are more likely over the age of 24, delaying postsecondary enrollment, not living on campus, being employed full time and attending school part time, opting for online and evening classes, having dependents other than a spouse, or being a single parent. We refer to them as nontraditional students.

The population of nontraditional students in the United States has been growing steadily for many years. They are quickly becoming the largest segment of the student population. About 74% of all 2011—12 undergraduate students had at least one nontraditional characteristic [3].

*These authors contributed equally to this work.

Nontraditional students often face significant challenges balancing academic pursuits with work and family commitments, which can hinder degree completion [4]. For example, eight years after starting their undergraduate studies in 2013, part-time students exhibited a higher transfer-out rate compared to full-time students. Specifically, the transfer-out rate for non-first-time part-time students was 29%, while it was 24% for first-time part-time students. Among full-time students, the transfer-out rate was 21% for first-time students and 18% for non-first-time students [5].

Despite these findings, our understanding of the specific factors impacting nontraditional student retention remains limited. It is unknown whether the knowledge of traditional student retention and mitigation measures can be applied to nontraditional students to achieve comparable outcomes.

To address this knowledge gap, higher education institutions require more robust data, analytical tools, and insights to support nontraditional students. Retention is critical not only for financial reasons but also as a key indicator of student success and well-being [6]. By gaining a deeper understanding of the challenges faced by nontraditional students, institutions can develop targeted interventions to improve retention rates and explain the decision-making process. This study aims to fill this knowledge gap by examining the factors influencing retention among nontraditional undergraduate students in a representative STEM major and building a predictive model to enable early interventions.

II. Related work

Retention in undergraduate Computer Science (CS) programs in the United States is an incredibly complex issue. Empirical data to examine retention are both limited and messy [7]. The Computer Science program is considered to be one of the most difficult programs because it requires a full comprehensive understanding and improved ability in order to fulfil academic requirements [8]. This may be why students seem to struggle to complete a CS degree. There are many works aiming to investigate student retention in CS education, including identifying variables related to gains of studying CS, the learning environment, degree's usefulness, and barriers as important predictors of students' intention to stay and complete their studies in CS [9].

There have been some conceptual research dedicated to nontraditional student retention, such as Bean's model of nontraditional undergraduate student attrition proposed in the mid-1980s [10]. The model highlights the importance of considering interactions between various factors (such as academic, background, psychological, and environmental variables) that can influence a student's decision to leave. Bean's model provided a valuable framework for understanding the complex factors influencing nontraditional student attrition. While the model itself might not be directly applicable today due to evolving educational landscapes, it laid the foundation for developing more nuanced retention strategies for nontraditional students.

More recently, some researchers found that nontraditional students persisted significantly more than their traditional counterparts in single courses and related assignments, but less in degree programs, creating a conundrum [11]. A follow-up study demonstrated that nontraditional students are more participatory than traditional students in a single course and are more engaged at deeper levels. Active learning activities increase the performance of nontraditional students, potentially increasing engagement and persistence. Moving toward higher levels of learning and engagement and lower levels of attrition can potentially resolve the conundrum [12].

In recent years, e-learning (online learning) has become very popular, especially during the COVID-19 pandemic. Compared to traditional classroom learning, e-learning has many benefits, such as higher flexibility, development of technical skills, continuous evaluation, and individual and collaborative activities [13]. Generally, education is made more accessible and affordable by e-learning, which is more desirable for nontraditional students. However, e-learning courses also result in higher dropout rates because distance education may create a sense of isolation in students who can feel disconnected from the other students, the instructors, and the university [14].

Educational research is advancing rapidly due to the vast amount of student data that can be used to create insightful patterns related to student learning. Educational Data Mining (EDM) uses the various amounts of data obtained from institutions to understand students' behaviors in educational institutions and to improve the teaching and learning environment [15]. A growing exploration area in EDM is predicting student academic performance [16]–[18]. For example, a recent study used algorithms such as Random Forests, Neural Network, Support Vector Machines, Logistic Regression, Naïve Bayes, and K-Nearest Neighbors to predict undergraduate student final exam grades [19]. They achieved a classification accuracy of 70–75%.

While many of these studies prioritize building models with high predictive accuracy, they often overlook the crucial aspect of interpretability. This lack of transparency makes it difficult to understand what factors the models actually use to make predictions, how the model arrives at a specific prediction, what additional information the model can potentially reveal, and the justifications behind its predictions. This lack of transparency is a growing concern in the age of explainable AI regulations, like the European Union's new regulations requiring that individuals affected by algorithmic decisions have a right to an explanation [20]. To ensure fairness and trust, we need models that are not just accurate, but also interpretable, allowing users to understand the reasoning behind the results.

III. Problem Statement

In this study, we investigate the nontraditional students who started as undergraduate Computer Science majors at a private, non-profit, teaching-intensive university in the Midwestern United States over a 13-year period. This open admission university specializes in nontraditional student education primarily delivered online. It offers five undergraduate majors in

computing technology: Computer Science, Cybersecurity, Information Systems, Information Technology, and Web Development. Computer Science has the largest enrollment. While the aggregated enrollment of these majors has been growing, our preliminary analytical observations [21] found that many students did not complete their programs, especially Computer Science.

The high number of nontraditional students at this university presents a unique opportunity to investigate their retention challenges. By addressing the following research questions, we aim to provide comprehensive answers that can inform effective strategies to improve their success:

1) **Research Question 1:** What are the primary factors influencing nontraditional undergraduate Computer Science student retention?
2) **Research Question 2:** How can we identify nontraditional undergraduate Computer Science students at a higher risk of dropping out at an earlier stage?

These research questions hold the key to unlocking better understanding of nontraditional undergraduate Computer Science student retention. Our research advancements will lead to more transparent machine learning models, enabling easier compliance with regulations and mitigating concerns about bias. Additionally, the models we build will provide academic leadership with powerful predictive capabilities. This will allow them to take targeted actions, such as enhancing student services and implementing early intervention programs, ultimately boosting student retention.

IV. Research Methodology

Our research leverages the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, a widely adopted methodology for data mining, analytics, and data science projects [22]. This robust and systematic framework ensures a comprehensive approach by guiding us through six key phases (Fig. 1). Each phase is clearly defined, preventing crucial steps from being overlooked.

A. Business Understanding

This is the first phase in the CRISP-DM framework. It focuses on understanding the objectives and requirements of the project. We have completed this phase in Section III by identifying the knowledge gap in nontraditional student retention and formulating the research questions for this study.

B. Data Understanding

The second phase involves identifying, gathering and exploring the preliminary data needed for our analysis. They include anonymized student information from undergraduate Computer Science majors enrolled at the university described in Section III. The data cover a period from Spring 2011 to Fall 2023, encompassing a total of 1,170 students.

The collected data contain information in the following categories:

- Demographic: Age, gender, and race
- Academic: GPA (including transfer, if applicable)

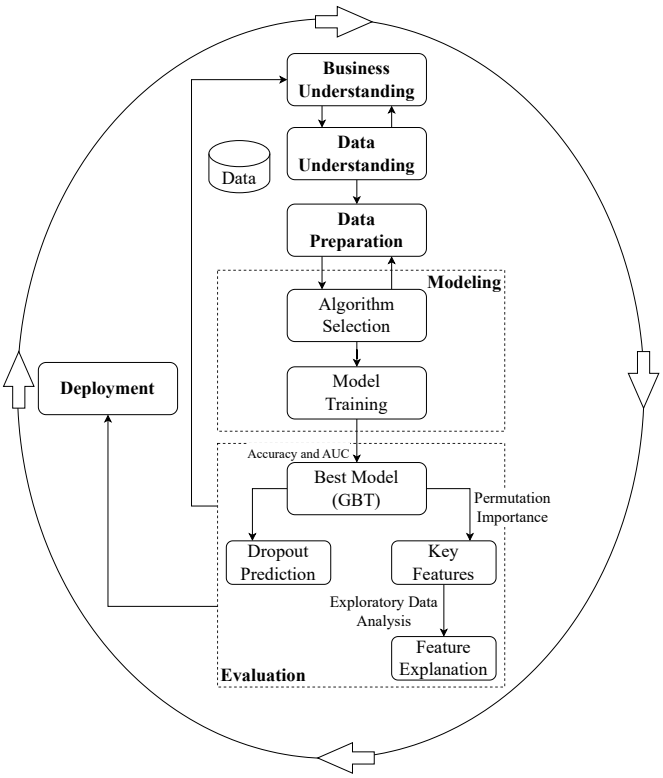


Fig. 1. Research Workflow

- Behavioral: Course load, course enrolled, course format (online/face-to-face), and prior college enrollment experience
- Financial: Financial aid information (awarded or not)
- Other: Mailing address, recruitment source

The field of Computer Science has undergone a substantial transformation in the past decade. The curriculum has adapted to keep pace with these advancements, incorporating new textbooks, instructional tools, assessment methods, and learning platforms. However, the core concepts and learning outcomes within the program have remained largely stable. Graduation and admission requirements have also seen minimal change. As a result, the longitudinal data collected for this study retain their validity.

C. Data Preparation

In this phase, we converted the raw data into the desired data format that can be analyzed in the modeling phase through data selecting, cleaning, constructing, integrating, and formatting.

The collected data do not contain all the features relevant to the definition of traditional and nontraditional students as described in Section I, such as dependents and status of employment. Therefore, we consider a student traditional if he or she satisfies the following two conditions:

- under the age of 25 in the first term of enrollment
- does not have any record of attending any other postsecondary institution after high school

Out of the total of 1,170 students, only 43 are traditional. Their data are discarded. The other 1,127 students are considered nontraditional. Unlike many other research works relying on data from traditional schools, which may include a limited number of nontraditional students, our study focuses exclusively on nontraditional students.

From a retention standpoint, students who switch majors are still considered retained by the university. In a separate study, we found a curricular factor impacting major switching [21]. Therefore, for the purpose of this research, we only count students whose claimed major is Computer Science in their last active term.

Measuring student retention can be challenging due to the lack of a universal definition and inconsistencies across institutions. This is particularly true for nontraditional students who may have unique enrollment patterns. In an earlier study, we observed that over 90% of students returned to their studies after a break of one to six terms [21]. Given the rarity of students returning after a longer absence, we can reasonably assume that students who have not enrolled for seven consecutive terms or more are unlikely to pursue their degrees at this university. Based on this rationale, we define a student's attrition label as `True` if they had not yet graduated and met one of the following two conditions:

- The student had officially reported withdrawal from the university.
- The gap between the student's last term of enrollment and Fall 2023 is more than six terms (equivalent to two years).

The attrition rate of students in the dataset is calculated as 57.3%.

Understanding the financial circumstances of students is important for analyzing retention. However, our initial data lack individual student income information. To address this limitation, we leveraged publicly available data on the latest median household income by zip code from the United States Census American Community Survey [23]. By incorporating these zip code-based estimates, we aim to gain insights into the potential influence of students' socioeconomic background on retention rates.

Through data preparation, we created 16 features for further analysis:

- Seven categorical features: Race, gender, financial aid information, student type (sources that they are recruited from), prior GPA (applicable to transfer students), online course enrollment history, and first-time college enrollment status
- Nine numerical features: Age in the first term, age in the last term, cumulative GPA, last term GPA, total credits taken, average credits per term, transfer credits (if applicable), percentage of credits taken face-to-face, and median household income

D. Modeling

In this phase, we evaluated five machine learning algorithms to predict student attrition: Logistic Regression, Support Vector Machines, K-Nearest Neighbors, Random Forest, and Gradient Boosting Trees. These models were chosen due to their strengths:

- Logistic Regression (LR): Offers interpretability, allowing us to understand the impact of different factors on retention.
- Support Vector Machines (SVM): Captures non-linear relationships between variables, potentially revealing complex patterns.
- K-Nearest Neighbors (KNN): Effective in identifying local patterns in the data, useful for understanding specific student subgroups.
- Random Forest (RF): Provides robustness and handles intricate interactions between features, leading to more accurate models.
- Gradient Boosting Trees (GBT): Improves prediction accuracy by sequentially refining the model based on previous errors.

To leverage the combined strengths of these approaches, we also created an ensemble model that incorporates all five algorithms. The models are trained and optimized by using scikit-learn [24].

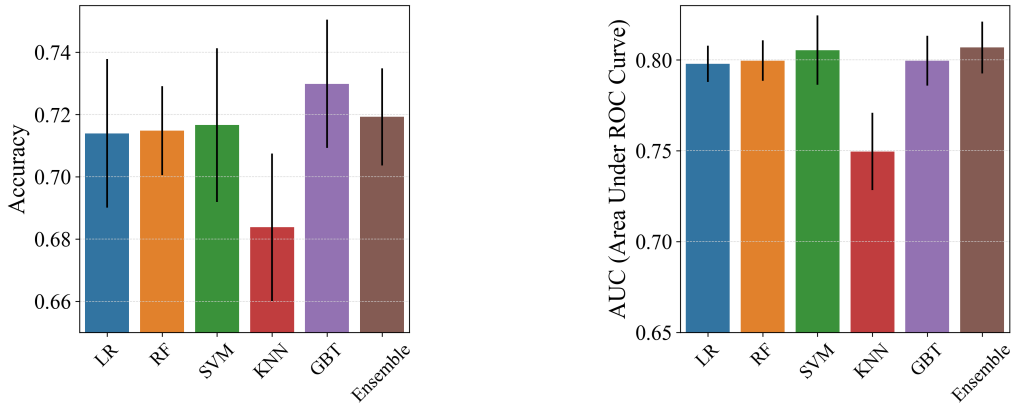
E. Evaluation

In this phase, we compared the predictive performance of our trained and optimized models to identify the model that most accurately predicts student attrition. To ensure a fair comparison and avoid overfitting, we employed a technique called nested cross-validation [25].

Nested cross-validation works by first dividing the dataset into two sets: outer folds and inner folds. Within each outer fold, we further split the data to create inner folds. Using these inner folds, we fine-tuned the hyperparameters of each model to optimize its performance on unseen data within that fold. After this tuning stage, we fit the model on the entire training data within the outer fold and tested it on the remaining unseen data (left-out data) from that fold. This process of inner fold tuning, model fitting, and testing on unseen data is repeated for all outer folds. To estimate the model's generalization error, which represents its performance on unseen data, we calculated the average test set score across all outer folds.

By separating data for hyperparameter tuning and performance evaluation, nested cross-validation offers several advantages. First, it minimizes data leakage. This means information from the tuning stage does not influence the evaluation stage, leading to a more accurate assessment of the model's ability to perform well on new data. Second, nested cross-validation helps reduce overfitting. This ensures the model generalizes well to unseen data, especially important when dealing with limited datasets. As a result of using nested cross-validation, our model evaluation is more robust and unbiased. This approach ultimately improves the model's ability to predict student attrition on new, unseen data.

To assess the effectiveness of our five optimized models in predicting student attrition, we employed two key metrics: accuracy and Area Under the ROC Curve (AUC).



(a) Accuracy Mean

(b) AUC Mean

Fig. 2. Model performance of Logistic Regression (LR), Random Forest (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Gradient Boosting Trees (GBT). Error bars represent the standard deviation.

Accuracy, a commonly used metric, represents the proportion of correct predictions made by the model. We calculated accuracy using a probability threshold of 0.5. In simpler terms, the model assigned a probability between 0 and 1 to each student indicating their likelihood of attrition. A threshold of 0.5 essentially divides these probabilities into two categories: above 0.5 suggests the student is "likely to leave" and below 0.5 suggests they are "likely to stay." Accuracy tells us how often this classification by the model matched the actual outcome (whether the student left or stayed enrolled).

AUC, or Area Under the ROC Curve, provides a more comprehensive picture. The ROC Curve itself plots the model's ability to correctly identify students who will leave (true positive rate) against the number of students incorrectly classified as leaving (false positive rate). This is calculated across various probability thresholds, meaning the model is not limited to the single 0.5 threshold used for accuracy. AUC summarizes the model's performance across all these thresholds into a single value between 0 and 1. A higher AUC indicates better overall performance in distinguishing between leaving and staying students.

By using both accuracy and AUC, we gain a balanced perspective on the models' performance. Accuracy offers a clear and direct measure, while AUC helps us understand the trade-off between correctly identifying leaving students and avoiding false alarms (classifying students who will stay as likely to leave). This combined approach leads to a more robust and nuanced evaluation of the models' effectiveness in predicting student attrition.

The model performance is compared in Fig. 2. The mean and standard deviation for both accuracy and AUC were obtained from the five iterations within the outer loop of the nested cross-validation process. Among the models, KNN has the worst performance, followed by LR. The improved results in nonlinear-kernel (SVM) hint at the presence of non-linear patterns in the data. SVM has good AUC measure, but its mean accuracy is worse than GBT. Tree-based models (GBT and RF) have better performance, indicating that the data may contain complex relationships between features and student attrition. These models are adept at handling feature interactions and are generally more resistant to overfitting.

Of all models, GBT stands out with the highest accuracy and an AUC very close to the highest. Its sequential building of trees allows for refined adjustments, possibly making it more attuned to the subtle factors influencing student retention. Therefore, we choose GBT as the preferred model for predicting students with higher risk of attrition.

As machine learning models become increasingly complex, the need for interpretability grows to ensure trust, accountability, and fairness. There are many approaches to model interpretability, such as global interpretability, local interpretability, model-specific interpretability, and model-agnostic interpretability [26]. We adopt feature importance to identify which data features contribute the most to predictions. This approach can help make sense of our model's predictions and improve the understanding of the model's overall behavior.

Several techniques exist to assess feature importance, and we opted for permutation importance [27]. This approach measures the decline in model performance when a single feature's values are randomly shuffled. A larger drop signifies the model's greater reliance on that feature for accurate predictions. Permutation importance holds several advantages: it is swift to calculate, widely used and understood, and aligns well with desirable properties for a feature importance metric [28]. Using permutation importance within our GBT model, we identified the top nine most impactful features (including six numerical and three categorical features) that exhibited an accuracy drop exceeding 0.9%. These features in Table I are considered to have the strongest influence on student retention.

V. Key Features

By delving deeper into these top features (Table I) through exploratory data analysis, we can glean further insights into their influence on student retention and elucidate the mech-

TABLE I
THE TOP NINE FEATURES ORDERED BY PERMUTATION IMPORTANCE IN GRADIENT BOOSTING TREES

Rank	Feature	Description	Accuracy Drop by Shuffled Feature
1	LastTermGPA	Student's Grade Point Average in the last term	5.2%
2	AvgCreditByTerm	Average number of credits a student takes per term	4.3%
3	FFCredPercent	Percentage of credits taken in face-to-face classes	4.1%
4	CreditsTaken	Total number of credits a student has taken	3.0%
5	CumulativeGPA	Student's overall Grade Point Average	1.9%
6	LastTermAge	Student's age in the last term	1.9%
7	StudentType_Community	Indicator for transfer student recruited through community college partnerships	1.7%
8	Online	Indicator for fully online student	1.0%
9	FinancialAid	Indicator for student receiving financial aid	0.9%

anisms behind these impacts. These key features can be used to establish a framework in which new knowledge can be discovered to enrich the comprehension of nontraditional undergraduate Computer Science student retention in its own context.

A. Academic Performance

Our dataset includes three features capturing student academic performance at different stages: *LastTermGPA* reflects a student's most recent academic performance, while *CumulativeGPA* captures their overall academic standing. *PriorGPA* is a unique feature of nontraditional students because they often (90% of the students in our dataset) transfer from another postsecondary institution. This is much less frequently observed in traditional students.

Consistent with prior research on traditional students [29], [30], our analysis reveals a strong positive correlation between academic performance and retention. Students with higher GPAs exhibit a lower probability of attrition. In both Fig. 3a and Fig. 3e, the `True` group includes students who dropped out of the university, while the `False` group includes students who did not drop out. We use the same notion in other numerical feature figures.

Among the three academic performance measures, we find that *LastTermGPA* holds the strongest influence (ranked first), followed by *CumulativeGPA* (ranked fifth). Interestingly, *PriorGPA* shows a weaker impact on retention and does not rank within the top nine features. This suggests that, for nontraditional students, recent academic performance may be a more sensitive indicator of their continuation decisions compared to their overall performance and performance at a previous institution. Overall, our findings reinforce the established notion that college grades serve as a key predictor of student persistence and degree completion.

B. Course Load

Research suggests a positive correlation between course load and student success. Studies have shown that students enrolling in more credits tend to achieve higher GPAs and experience greater retention rates [31], [32]. Our analysis aligns with these findings, revealing a strong impact of course load on student persistence. The features *AvgCreditByTerm* (ranked second) and *CreditsTaken* (ranked fourth) highlight this connection. As Fig. 3b and Fig. 3d illustrate, students taking more credits per term or cumulatively exhibited a lower probability of attrition (the `False` group).

It is important to note that course load selection can be influenced by academic ability and prior academic achievements. Students who choose to take on a heavier course load might demonstrate a stronger focus on their studies, which could contribute to improved retention.

C. Class Format

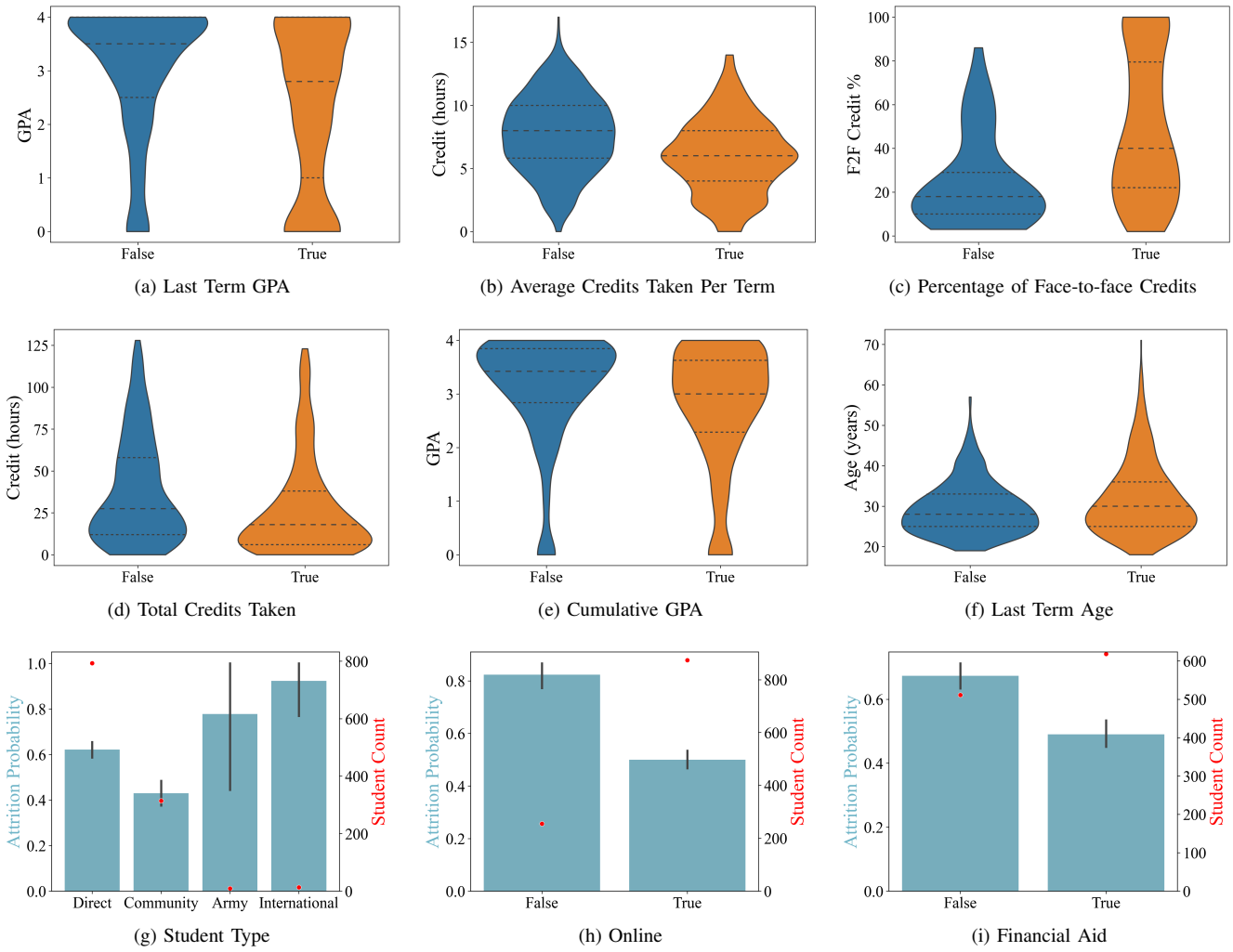
While some research suggests higher attrition rates for online students [33], [34], our analysis reveals a contrasting trend. Fig. 3h demonstrates that the majority of students (872, or 77.4% of the total number) who took classes entirely online (the `True` group) exhibited a significantly lower probability of attrition (0.5) compared to those who took one or more face-to-face (FF) classes (0.82).

To further explore this, Fig. 3c examines the distribution of students based on the percentage of FF classes taken. Since most students did not take any FF classes, the figure focuses on students with a positive FF percentage. It is evident that students taking a higher proportion of FF classes have a greater chance of dropping out (the `True` group). This observation aligns with the strong impact of the *FFCreditPercent* feature, ranked third in Table I.

In essence, Fig. 3c and 3h highlight a key characteristic of our nontraditional student population: they tend to thrive in online learning environments compared to traditional students who may benefit more from face-to-face interaction.

D. Age

Age serves as a proxy for the heterogeneous nature of the nontraditional student population. While traditional student demographics at private, non-profit four-year institutions skew younger, with 87% of full-time undergraduates under 25 in Fall 2021 [35], our study reveals a significantly older population. In our dataset, 50% of students are 29 years or older in their last active term, with an average age of 30.8. Unsurprisingly, Fig. 3f demonstrates a correlation between student's age in the last term (the *LastTermAge* feature) and attrition, with older students exhibiting a higher probability of dropping out (the `True` group). This can be attributed to the additional challenges faced by adult learners, particularly those juggling family, work, and other life commitments alongside academic pursuits.



(a) Last Term GPA (b) Average Credits Taken Per Term (c) Percentage of Face-to-face Credits

(d) Total Credits Taken (e) Cumulative GPA (f) Last Term Age

(g) Student Type (h) Online (i) Financial Aid

Fig. 3. Student attrition probability associated with the top nine features. Numerical features (a–f) are plotted as violin charts where the middle lines represent the median, the top lines denote the 75th percentile, and the bottom lines indicate the 25th percentile. `True` groups include unretained students. The width of a violin plot represents the density of data points at a particular value on the y-axis. Categorical features (g–i) are plotted as bar charts where red dots represent student counts, blue bars represent attrition probabilities, and error bars represent the 95% bootstrap confidence interval. The `True` group in (h) includes students taking all credits online; the `True` group in (i) includes students receiving financial aid.

Interestingly, the age distribution in both retained (`False`) and unretained (`True`) groups leans towards older demographics. The median last term age of the unretained group is 30, slightly older than the retained group (28). There are very few students under 25 in both groups, a stark contrast to the typical profile of a traditional student.

E. Recruitment Source

Nontraditional students are more likely to transfer between institutions of various types. Research suggests two main groups of transfer students: those with a pre-defined transfer goal (e.g., medical school) and those whose plans change or lack a clear direction [36], [37]. In our study, most students fell into the latter category. 1,010 out of the total 1,127 students transferred credits from another school. That is 90% of the population. This university also boasts a diverse recruitment strategy, attracting students beyond direct applications. Collaborations with the military, corporations, and especially community colleges are key recruitment channels. A designated *StudentType* feature tracks these sources.

There are 314 students (27.9% of the entire population) recruited through community college partnerships. They are the `Community` student type in Fig. 3g. Interestingly, these students have the lowest attrition probability (0.43). This suggests a stronger effect from the partnership compared to simply transferring from a community college.

The university's strong relationship with community college partners likely contributes to this success. Initiatives like a 3+1 transfer program (three years at a community college, plus one year at the university), a user-friendly online transfer tool, and reduced tuition for eligible students further support this collaboration. Additionally, curricular articulation agreements, particularly within the state, ensure smooth course transfer while maintaining academic rigor.

Our findings highlight the crucial role of two-year institutions in preparing nontraditional students for bachelor's degrees. Fostering closer relationships between four-year and two-year institutions through strategic partnerships and curricular alignment can significantly improve student retention.

F. Financial Aid

Unlike traditional students who often rely on parental support, nontraditional students are typically financially independent. While this independence fosters self-reliance, it can also create financial burdens that threaten their academic journey. Our study reinforces the positive impact of financial aid on nontraditional student retention as evidenced by the *FinancialAid* feature, which is shown in Fig. 3i. Students with some form of financial aid (617 in the `True` group) exhibited a significantly lower attrition probability (0.49) compared to those solely relying on personal finances (510 students in the `False` group with a 0.67 attrition probability).

The benefit of financial aid for nontraditional students extends beyond simply bridging the financial gap. The specific types of aid available can significantly impact their ability to persist. Unlike traditional students who may primarily rely on federal grants and loans, nontraditional students benefit from a wider range of financial aid options. These include federal and state grants, scholarships, and employer tuition assistance. For many employed nontraditional students, programs offered by their employers for tuition reimbursement can be a game-changer. This financial support fosters a win-win situation, as it promotes employee retention and skill development.

To summarize, our analysis identified nine key features influencing nontraditional undergraduate Computer Science student retention using a GBT model's feature permutation importance. These features can be grouped into categories like academic performance, course load, class format (online/face-to-face), demographics (age), recruitment source, and financial aid. By analyzing these categories, we discovered unique retention patterns specific to nontraditional students compared to traditional students.

While some factors, like academic performance, age, and finances, affect both student groups similarly, the strength of their influence might differ. Further research is necessary to fully understand these nuances.

VI. Conclusions and Future Work

Nontraditional students comprise the majority of the undergraduate population in the United States today. While facing more challenges to complete college, especially in STEM fields like Computer Science, their unique retention factors remain largely unexplored. The evolving student demographic demands innovative strategies for student success. To effectively support nontraditional students, a deeper understanding of their specific needs, particularly regarding retention, is crucial. This research employs the CRISP-DM framework to investigate nontraditional undergraduate Computer Science student retention at a university specializing in this population. We aim to answer two key research questions:

1. **Identifying the key features impacting nontraditional undergraduate Computer Science student retention**. Our analysis suggests that academic performance, class format, course load, age, recruitment source, and financial aid availability are the major key features impacting retention. Unveiling these key features and how they impact retention not only aids retention efforts but also enhances the interpretability of machine learning models, promoting ethical considerations and mitigating potential bias. This knowledge foundation allows for further exploration and a more nuanced understanding of the specific context influencing nontraditional student retention.

2. **Developing predictive models to identify students who have high risk of attrition**. Among the five machine learning models we trained and evaluated, Gradient Boosting Trees (GBT) emerged as the most effective in predicting student attrition. This success is attributed to our nested approach to model training and optimization, which helps to minimize data leakage and overfitting. This results in a more robust and unbiased model, ensuring its generalizability to new, unseen data. This is a crucial advantage when dealing with datasets of limited size. Our GBT model demonstrates strong accuracy in identifying students at high risk of attrition, which can empower academic leadership to make proactive decisions regarding intervention strategies.

The final stage of the CRISP-DM framework, Deployment, focuses on translating our findings into actionable strategies. We plan to collaborate with the university's student service department to develop targeted interventions, including leveraging the GBT model's predictions to identify students at risk of attrition and design personalized support programs to address their specific needs.

This research provides a roadmap for targeted interventions that can potentially benefit many other colleges with similar nontraditional student demographics. Student support services may be tailored according to the identified impacting features. For instance, early academic interventions for students with lower GPAs, strong partnership with community colleges, enhanced online instruction, and expanded financial aid opportunities could be implemented. Additionally, the machine learning model's predictive capabilities can be used to proactively identify students at risk of dropping out, allowing for timely interventions such as academic advising, tutoring, or mentorship programs. Ultimately, by understanding the unique challenges faced by nontraditional CS students, institutions can develop tailored support services to enhance their academic success and persistence.

Building on the success of this work, we plan to extend our research in two key ways:

- Include more academic disciplines: By analyzing data from additional majors, we can gain a broader understanding of factors affecting nontraditional student retention across different programs.

- Focus on first-to-second-term retention: This deeper dive will provide valuable insights into the critical transition period between the first and second semesters, a crucial stage for student retention.

REFERENCES

[1] X. Chen, "Stem attrition: College students' paths into and out of stem fields. statistical analysis report. nces 2014-001." *National Center for Education Statistics*, 2013.

[2] National Center for Education Statistics. (2024) Postsecondary outcomes for nontraditional and traditional undergraduate students. condition of education. [Online]. Available: https://nces.ed.gov/programs/coe/indicator/ctu

[3] A. W. Radford, M. Cominole, and P. Skomsvold, "Demographic and enrollment characteristics of nontraditional undergraduates: 2011-12. nces 2015-025." *National Center for Education Statistics*, 2015.

[4] G. Markle, "Factors influencing persistence among nontraditional university students," *Adult Education Quarterly*, vol. 65, no. 3, pp. 267–285, 2015. [Online]. Available: https://doi.org/10.1177/0741713615583085

[5] V. Irwin, K. Wang, T. Tezil, J. Zhang, A. Filbey, J. Jung, F. B. Mann, R. Dilig, and S. Parker, "Report on the condition of education 2023. nces 2023-144." *National Center for Education Statistics*, 2023.

[6] Watermark Insights. (2022) Why is college retention important? [Online]. Available: https://www.watermarkinsights.com/resources/blog/retention-vs-enrollment

[7] C. Stephenson, A. D. Miller, C. Alvarado, L. Barker, V. Barr, T. Camp, C. Frieze, C. Lewis, E. C. Mindell, L. Limbird *et al.*, *Retention in computer science undergraduate programs in the us: Data challenges and promising interventions*. ACM, 2018.

[8] H.-T. Wu, P.-C. Hsu, C.-Y. Lee, H.-J. Wang, and C.-K. Sun, "The impact of supplementary hands-on practice on learning in introductory computer science course for freshmen," *Computers & Education*, vol. 70, pp. 1–8, 2014.

[9] M. N. Giannakos, I. O. Pappas, L. Jaccheri, and D. G. Sampson, "Understanding student retention in computer science education: The role of environment, gains, barriers and usefulness," *Education and Information Technologies*, vol. 22, pp. 2365–2382, 2017.

[10] J. P. Bean and B. S. Metzner, "A conceptual model of nontraditional undergraduate student attrition," *Review of educational Research*, vol. 55, no. 4, pp. 485–540, 1985.

[11] H. Ellis, "A nontraditional conundrum: The dilemma of nontraditional student attrition in higher education," *College Student Journal*, vol. 53, no. 1, pp. 24–32, 2019.

[12] ——, "Pursuing the conundrum of nontraditional student attrition and persistence: A follow-up study," *College Student Journal*, vol. 53, no. 4, pp. 439–449, 2020.

[13] T. Daradoumis, A. A. Juan, F. Lera-López, and J. Faulin, "Using collaboration strategies to support the monitoring of online collaborative learning activity," in *Technology Enhanced Learning. Quality of Teaching and Educational Reform*, M. D. Lytras, P. Ordonez De Pablos, D. Avison, J. Sipior, Q. Jin, W. Leal, L. Uden, M. Thomas, S. Cervai, and D. Horner, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 271–277.

[14] L. Calvet and A. Juan, "Educational data mining and learning analytics: differences, similarities, and time evolution," *RUSC. Universities and Knowledge Society Journal*, vol. 12, p. 98, 07 2015.

[15] J. M. Ortiz-Lozano, A. Rua-Vieites, P. Bilbao-Calabuig, and M. Casadesús-Fa, "University student retention: Best time and data to identify undergraduate students at risk of dropout," *Innovations in education and teaching international*, 2018.

[16] C. Jalota, "An effectual model for early prediction of academic performance using ensemble classification," *Journal of Language and Linguistics in Society (JLLS) ISSN 2815-0961*, vol. 3, no. 02, pp. 19–33, 2023.

[17] P. Ghosh, R. Roy, S. Mandal, M. M. Chowdhary, and S. Bokshi, "Data mining approach to predict academic performance of students," *BOHR International Journal of Computer Science*, vol. 2, no. 1, pp. 59–69, 2023.

[18] S. M. Dol and P. Jawandhiya, "A review of data mining in education sector," *Journal of Engineering Education Transformations*, vol. 36, no. Special Issue 2, 2023.

[19] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, 2022.

[20] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[21] C. Chu and J. Li, "Learning analytics finds that a shared course may improve technology students retention," *J. Comput. Sci. Coll.*, vol. 38, no. 6, p. 64–71, jun 2023.

[22] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. Manchester, 2000, pp. 29–39.

[23] The United States Census Bureau. (2021, Mar.) Median household income. [Online]. Available: https://data.census.gov/

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC bioinformatics*, vol. 7, pp. 1–8, 2006.

[26] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book

[27] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010.

[28] D. Becker and A. Cook. (2023, Mar.) Permutation importance. [Online]. Available: https://www.kaggle.com/code/dansbecker/permutation-importance

[29] D. D. Gifford, J. Briceno-Perriott, and F. Mianzo, "Locus of control: Academic achievement and retention in a sample of university first-year students." *Journal of college admission*, vol. 191, pp. 18–25, 2006.

[30] R. D. Reason, "Student variables that predict retention: Recent research and new developments," *NASPA journal*, vol. 46, no. 3, pp. 482–501, 2009.

[31] R. F. Szafran, "The effect of academic load on success for new college students: Is lighter better?" *Research in Higher Education*, vol. 42, pp. 27–50, 2001.

[32] N. Huntington-Klein and A. Gill, "Semester course load and student performance," *Research in higher education*, vol. 62, no. 5, pp. 623–650, 2021.

[33] R. F. Kizilcec and S. Halawa, "Attrition and achievement gaps in online learning," in *Proceedings of the second (2015) ACM conference on learning @ scale*, 2015, pp. 57–66.

[34] M. Shaw, S. Burrus, and K. Ferguson, "Factors that influence student attrition in online courses," *Online Journal of Distance Learning Administration*, vol. 19, no. 3, pp. 211–231, 2016.

[35] V. Irwin, J. De La Rosa, K. Wang, S. Hein, J. Zhang, R. Burr, A. Roberts, A. Barmer, F. Bullock Mann, R. Dilig *et al.*, "Report on the condition of education 2022. nces 2022-144." *National Center for Education Statistics*, 2022.

[36] L. Aulck and J. West, "Attrition and performance of community college transfers," *PloS one*, vol. 12, no. 4, p. e0174683, 2017.

[37] M. Blekic, R. Carpenter, and Y. Cao, "Continuing and transfer students: Exploring retention and second-year success," *Journal of College Student Retention: Research, Theory & Practice*, vol. 22, no. 1, pp. 71–98, 2020.