Fighting College Attrition Through Timely Interventions: A Sequential Machine Learning Model

Juan Carlos Apitz
California State University Long Beach
juan.apitz@csulb.edu

Mahmoud Albawaneh
California State University Long Beach
mahmoud.albawaneh@csulb.edu

Janaki Santhiveeran
California State University Long Beach
janaki.s@csulb.edu

Dhushy Sathianathan
California State University Long Beach
dhushy.sathianathan@csulb.edu

This is the abstract. It should contain from 100 to 250 words.

Keywords: attrition, machine learning, CatBoost, XGBoost

1. Introduction

- TOO US CENTRIC (WHICH FEDERAL GOVERNMENT)
- CONTEXT SHOULD INCLUDE CSU, INITIATIVE 2025
- Attrition is a significant problem for universities.
- Students leave for a variety of reasons. Some students transfers to other institutions but the majority simply leave without obtaining a degree.
- In this paper we use machine learning to predict and fight attrition.

The federal government spends $234.9 billion on college education by providing student loans and grants (The College Board, 2021), as national policies focus on making higher education accessible to everyone. As college education costs skyrocket, more than 50% of undergraduate students graduate in debt (The College Board, 2021). Consequently, the main goal of higher education institutions is shifting its emphasis to retention from access toward reducing dropouts. Student attrition is the most complicated problem in the university system (Kim & Kim, 2018). Fewer than two-thirds of students graduate within six years (U.S. Department of Education, 2021), and graduation numbers are much lower for Black (42%) and Hispanic (56%) students nationally (Musu-Gillette et al., 2018). Despite numerous efforts to soften its effect, attrition remains one of the most severe issues confronted by four-year universities. One of the troubling facts about attrition is that it occurs each semester.

Centered on the idiosyncratic disadvantages of attrition, researchers, policymakers, and educators consider this issue a considerable burden to the university system (Kim & Kim, 2018). Hence, attrition negatively affects the economy, contributing to low gain on investment and waste of taxpayer money on higher education.

College attrition means a student not enrolling in a semester to stay at a university. This research focuses on attrition during s_2, s_3, and s_4 based on the students who did not enroll in courses. Since most students voluntarily attrition in the beginning year(s) of their undergraduate college education (Bargmann, Thiele, & Kauffeld, 2022), this study focuses on the first three semesters of undergraduate education. Such definition is limited as it includes students who transfer to four-year colleges or due to academic dismissal.

Several factors contribute to college attrition. Estimating the future attrition risk in multiple domains, such as personal characteristics, academic preparation (Bishop & Bailey, 2021; Bargmann, Thiele, & Kauffeld, 2022), academic performance, and educational cost (Bishop & Bailey, 2021), were explored mainly using survey research. However, there is limited research on these factors, and an assortment of predictors should be construed in combination and not in seclusion to help in risk estimation. To foresee an episode of potential attrition, it is essential to distinguish students who attrition due to low academic performance, a precursor to attrition. The exact probability model must be constructed on analytical factors integrated into forecasting models using ML algorithms. Such analytical models could be instrumental for detecting at-risk college students with a greater probability of attrition in several domains to assess individual profiles that suggest impending problems in particular attrition realms.

Data are an indispensable part of ML. Due to increased student data, there has been growing attention to student data and ML modeling. Prediction simulations are valuable in designing prevention strategies to prevent or minimize attrition. Although complicated, it is necessary to develop ML models to understand and produce predictions regarding the factors contributing to attrition to establish effective control and long-term or short-term prevention strategies and policies. Predicting student attrition accurately could help alleviate its social and economic costs (Robison et al., 2017).

This study adds to existing research in various ways. First, this study focuses on prediction by using ML algorithms to find patterns in often rich and unwieldy data. Thus, we added several algorithms, including LR, RF, CatBoost, XGB, and LGBM to identify the best possible ML models for predicting attrition. Hence, this study extends the understanding of the application and efficacy of ML models in higher education. This study aims to examine which ML models accurately predict attrition. Second, this study expanded the prior research using predictive analytics that will account for various predictors (e.g., demographics, pre-entry data, academic performance, and academic goal engagement) by combining data from the first three semesters. To identify students at risk of attrition in the future, precise risk estimation must be based on pertinent analytical factors integrated into risk prediction models. Finally, this study identified the most critical features in explaining student attrition. Analyzing and predicting the reasons for attrition can form a basis for advisement plans, intervention policies, targeted advising, and developing prevention strategies for the university system, advisors, and students. The ML algorithms' performances will help us to design web applications to predict students who will attrition in s_4 using s_1, s_2, and s_3 performance data. Such information will help with early intervention and advising at-risk students before dropping out of university.

2. Related Work

The literature survey in educational data analysis is elaborated on in this section. The proposed research is based on the earlier theory of student integration (Tinto, 2006), which suggests that students' dropout is due to their past and current academic performance. Others believe that student dropout is due to their lack of focus on academic performance and goal engagement (Álvarez-Pérez et al., 2021). Undecided students tend to drop out or have low GPAs than those with clear educational goals (Pickenpaugh et al., 2022; Swanson, Vaughan, & Wilkinson (2017). Several other factors, such as distance between home and school, social and emotional connection, ethnicity, and first-generation status, contribute to college dropout (Álvarez-Pérez et al., 2021; Pickenpaugh et al., 2022). Some explored roles of financial aid in determining college dropout (Lee et al., 2021). Existing research on attrition is often survey-driven, surveying students or programs (Álvarez-Pérez et al., 2021). However, the proposed study is data-driven, an analytical approach using enrollment data from a four-year state university, California State University, Long Beach.

ML is an assuring method for constructing a prognostic model for dropouts and offers early notice to responsible advisors to take preventive measures to help individual students at the risk of dropping out (Bonifro et al., 2020). In recent years, growing number of researchers used ML classification methods such as LR (Aulck et al., 2016; Aulck, et al., 2019; Chen, Johri, & Rangwala, 2018; Kemper, 2020), XGB (Aulck et al., 2019; Kiss et al., 2021; Moghimi & Metzger, 2022; Moreira de Silva et al, 2022), RF (Aulck et al., 2016; Moreira de Silva et al, 2022), NN (Kiss et al., 2021; Moghimi & Metzger, 2022; Moreira de Silva et al, 2022), SVM (Moghimi & Metzger, 2022;), DT (Kemper, 2020; Moghimi & Metzger, 2022) and other algorithms (Aulck et al., 2016; Kiss et al., 2021; Moghimi & Metzger, 2022; Aulck et al., 2016) to form predictive models. Other researchers have used ensemble learning and machine learning algorithms (Kemper, 2020) to improve accuracy.

Several studies have employed machine learning algorithms to identify attrition patterns and dropout rates using data from various universities in the USA (Aulck et al., 2016; Aulck et al., 2019; Delen, Topus, & Eryarsoy, 2020; Moghimi & Metzger, 2022). Studies from the USA have reported contrasting results in prediction accuracy. For example, in a survey of undergraduate students between 1998 and 2006 from the University of Washington (UW), Aulck et al. (2016) used LR, RF, and KNN to predict dropout. LR emerged as the best model in predicting dropout accurately at 66.59

In a study of 35,021 American citizens from 2006 to 2015 from a University in the USA using BNN, Delen, Topus, and Eryarsoy (2020) had 20

Several researchers from abroad, including Germany (Kemper, 2020), Hungary (Kiss et al., 2021), China (Niyogisubizo et al., 2022), and Portuguese (Moreira de Silva et al., 2022), have effectively implemented ML algorithms to predict dropout. For instance, Kemper (2020) forecasted student dropout using 2,556 Industrial Engineering students from Germany who graduated and 620 students who had dropped out of Karlsruhe Institute of Technology (KIT) between 2007 and 2012 using LR and DT. Judging by the sensitivity, both LR and DT predicted dropouts better using balanced (86.2

Moreira de Silva et al. (2022) used only academic grades in a study of student dropouts at a Portugal university. Moreira de Silva et al. (2022) used RF, XGB, CatBoost, and ANN on the final test set. They concluded that XGB showed the best result, with the XGB model predicting nearly 92

2.1. Conclusion

Studies presented in this section used data from enrolled students to develop models to predict mostly dropouts. Some used training sets (Kemper, 2020; Moghimi & Metzger, 2022) and others used test sets (Aulck et al., 2016; Aulck, et al., 2019; Moghimi & Metzger, 2022; Moreira de Silva et al, 2022; Niyogisubizo et al., 2022). Some researchers used unbalanced (For example, Aulck et al., 2019; Delen, Topus, and Eryarsoy, 2020; Kemper, 2020; Moreira de Silva et al., 2022) and balanced datasets (E.g., Delen, Topus, and Eryarsoy, 2020; Kemper, 2020) datasets. These studies give a basis for using imbalanced datasets in our ML models.

Among the ML algorithms reviewed in this section, LR (Aulck et al., 2016; Aulck et al., 2019; Chen, Johri, & Rangwala, 2018, Kemper, 2020) and XGB (Aulck et al., 2019; Kiss et al., 2021; Moghimi & Metzger, 2022; Moreira de Silva et al., 2022) emerged as the best performing models. RF (Moreira de Silva et al., 2022) and stacking ensemble (Niyogisubizo et al., 2022) performed significantly better in a few studies. Consequently, the researchers chose mainly these models to predict dropout, and the models were evaluated using the following metrics.

Several types of evaluation metrics, such as accuracy, recall, sensitivity, and precision (Aulck et al., 2019; Delen, Topus, and Eryarsoy, 2020; Kemper, 2020; Niyogisubizo et al., 2022) were reviewed in this literature review. Some researchers used AUC for the ROC curve score (Aulck et al., 2016; Aulck et al., 2019; Chen, Johri, & Rangwala, 2018; Delen, Topus, & Eryarsoy, 2020) in unbalanced datasets. Among the best performing models reported by the researchers, LR and ADA models reported the highest accuracy of 99.3

The ML models reviewed in this section used several features in their predictive models. Some used demographic data (Delen, Topus, & Eryarsoy, 2020; Kiss et al., 2021; Moreira de Silva et al., 2022), academic performance (Aulck et al., 2019; Delen, Topus, & Eryarsoy, 2020; Kiss et al., 2021; Moghimi & Metzger, 2022; Moreira de Silva et al., 2022), and goal engagement (Moghimi & Metzger, 2022) features as predictors of dropout. These studies guided in selecting study features.

3. Student Population and Datasets

The population of study in this project consists of CSULB First Year Students (FYS) at the undergraduate level who begin classes during the Fall of each academic year (Fall cohorts). To conduct the analysis we obtained a data set consisting of student records for the Fall cohorts from 2018 to 2020.

3.1. Data Features and Feature Engineering

Student features can be classified as:

- Demographic Characteristics
- Pre-entry Academic Performance
- Semester (s_i) Academic Performance

Examples of a figure and a table are given in Figure 1 and Table 1.

Table 1: This is an example of a table that lists the margins of this template. Captions should follow the same rules as a figure, except that they are put on top of the table.

Category	**Variable Label**	**Description**	**Type**
Pre-entry Variables	HS Overall GPA	Overall HS GPA at entry	Numerical
	HS Math GPA	HS GPA in Math classes	Numerical
	HS English GPA	HS GPA in English classes	Numerical
	Load Index Pre-entry	Percentage of units transferred	Engineered

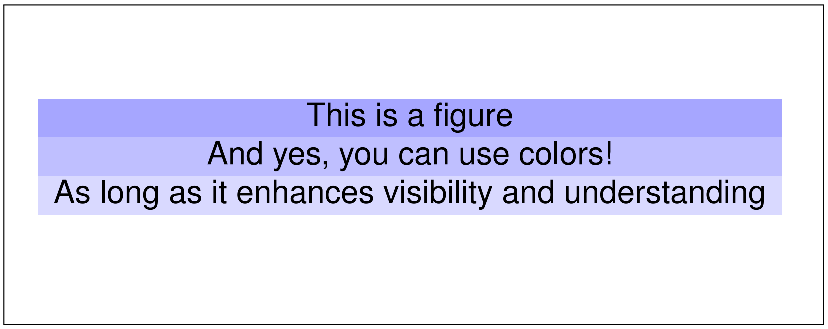
This is a figure
And yes, you can use colors!
As long as it enhances visibility and understanding

Figure 1: This is the figure's caption. It should be a centered paragraph of width 193mm (7.6in). Font size should be 11pt.

4. Appendices

Supplementary technical material (e.g., mathematical proofs or descriptions of experimental procedures) should be collected in an appendix at the end of the paper (before the acknowledgements and the references sections).

5. Footnotes and acknowledgments

Footnotes should be used sparingly and indicated by consecutive superscript numbers in the text. Material to be footnoted should appear at the bottom of the page on which it is referenced. Acknowledgments and grant numbers should be put into a separate 'Acknowledgment' section right before the list of references.

6. References

References should follow the ACM standard. The example provided here uses the `jedm.cls` class file and `acmtrans.bst` bib style file. For example, we could write that **?**) published a review on EDM in this journal; other reviews were published later (**?**, for eg.). The provided ref.bib file contains examples of virtually every possible citation type.

7. Supporting materials

Authors are encouraged to submit the data they use and the analysis code in order to replicate and perform rigorous comparisons across studies. The data and code can be stored on the educationaldatamining.org site for public reference, or stored privately for reviewing purpose only if required. See the online submission instructions for guidance on using and citing code repositories.

8. Page numbering and sectioning

For the manuscript to be reviewed, page number should appear at the bottom of each page. **For the final version, they must be taken out as the standard JEDM footer is be added.**

8.1. Sections and subsections

Section style should follow the example in this document.

8.1.1. Subsection levels

There should be no more than three levels of sections.

Fourth level. The fourth level should simply have their title in the same font style as those of of subsections, without numbering and at the beginning of a paragraph.

9. Submission

The journal prefers PDF format, but can also accommodate most other popular formats. Care should be taken to embed fonts in the rendered PDF. The provided Word template has a larger filesize because it contains embedded fonts.

Instructions for submitting the papers on the website are at http://jedm.educationaldatamining.org/index.php/JEDM/about/submissions#onlineSubmissions. You first need to register and log in to the site. When registering, you must activate your role as “author”.