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

Mahmoud Albawaneh and Juan Carlos Apitz

Chapter 1

Introduction

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 s4 using s1, s2, and s3 performance data. Such information will help with early intervention and advising at-risk students before dropping out of university.

Retention rate is a standard metric used universally by higher education institutions in the United States. It measures institutions ability to retain students over time once students enrolled and started their education journey for the first time at the beginning of their first term or semester.

For 4-year institutions and above, the first-year retention rate is the percentage of first-time or transfer degree-seeking undergraduates who did not persist or did not come back a year later (Began in January 2017 and did not persist or did not come back in January 2018). Therefore, the first-year retention rate ranges between 0

The first-year retention rate is widely used in higher education and it is the most popular retention rate metric. In addition, second-year, third-year and fourth-year retention rates are commonly used as retention performance metrics for higher learning institutions. Data and reports on retention rates are published and publicly available on every institution website as well as the institutional research offices websites. Retention rates are accessible and available on the official institutions disclosures, annual reports, institutional and programmatic accreditation reports, and various federal, state and private agencies publications. IPEDS, which is part of the federal department of education, mandates each institution to report first-year retention rates annually due to its importance and significant impact on other performance and academic metrics. Retention rate generally affects continuing student's enrollment, new enrollment goals, graduation rate and institutional budget.

In this step-by-step guide, you will learn how to build classification model utilizing logistic regression to predict which students will enroll or not enroll a year later.

Upon completion of this guide, you will be able to:

- Understand and explore retention data set
- Prepare retention dataset for machine learning
- Split prepared retention dataset into training and testing datasets
- Choose and use proper evaluation metric for machine learning
- Train and fit logistic regression model
- Operationalize logistic regression model for new and unseen data

1.1 Guide Overview

This guide is divided into the following sections:

1. Machine Learning Process
2. Retention dataset as practical case study
3. Exploration and visualization of retention dataset
4. Preparation of retention dataset to become machine-learn ready
5. Model building, training and testing
6. Model deployment

1.2 Machine Learning Process

Building a machine-learning model from start to finish is a seven steps process. It is crucial that this process is followed sequentially step by step without skipping, avoiding, ignoring or discarding any step. Following the seven steps process properly and diligently from step one through step seven ensures production of an accurate and reliable machine-learning model that can be put in use in production environment.

The seven steps of the machine learning process are:

1. **Define problem statement:**
   It is important to define the problem or the challenge with explicit objectives and proper scope. For the retention rate dataset case study, the primary objective is to build a classification model in order to predict which students will enroll or not enroll in the following fall semester who previously enrolled in the previous fall semester. The emphasis is on prediction not on model interpretation or explanation. Therefore, the strategy is to achieve a high value for the accuracy or any other appropriate chosen metric regardless of the machine learning model complexity and explanatory power.

2. **Extract or Collect data**
   In higher education institutions, data is usually available in traditional databases or data warehouses. Based on the problem statement definition, the required dataset is queried from different tables within the database or data warehouse then merged using common institutional identifier. This represents an example of direct extraction of the required dataset from institutional database. On the contrary, the required dataset may not exist in the institutional databases or data warehouses. In this particular case, data will need to be collected using appropriate data collection methods such as institutional surveys.

3. **Prepare data**
   Once the required dataset is either extracted or collected, data must be prepared properly in order to be digested and processed by the machine-learning models. The basic tasks of data preparation includes:

   - Scaling or normalizing the numerical data
   - Converting string or text data into numerical data
   - Removing or imputing missing values

- Handling imbalanced dataset (i.e. retention rate dataset is example of imbalanced dataset)

4. **Split data into training and testing**
   It is critical to split the prepared dataset into training and testing dataset. The common rule for splitting the prepared dataset is the 70%/30% rule. This means 70% of the entire prepared dataset is allocated for training and the other 30% is used for testing. The splitting rule is a hybrid of art and science and it depends primarily on the amount of available data despite the popularity of the 70%/30% common rule. A common mistake is to use the testing dataset for data exploration, correlation and visualization. The testing dataset must be put aside and never used except for model evaluation.

5. **Choose a machine learning model**
   There are numerous machine-learning models to choose from varying in model complexity and the number of hyper-parameters. Hence choosing the best model can be tricky and may require trying different models. Choosing a machine-learning model depends on the problem statement definition. If the primary goal is prediction not interpretation then the model producing the highest value for the evaluation metric should be chosen. On other hand, if the primary goal is interpretation of the findings then a simple and highly interpretable model should be selected.

6. **Fit the chosen machine learning model**
   Once the dataset is properly prepared and a machine-learning model is selected, the model building begins by performing the following three essential tasks:

   - Fit or fine-tune the machine-learning model using the training dataset. Fine-tuning the machine-learning model is finding the best or optimum hyper-parameters.
   - Evaluate the fitted or fine-tuned machine-learning model using the testing dataset.
   - Select the fitted or fine-tuned machine-learning model producing the highest value for the chosen evaluation metric outlined in step one

7. **Deploy the fitted or the fine-tuned machine learning model**
   At this stage, the fine-tuned machine-learning model is ready for deployment in production environment. This is where numerous machine-learning projects fail and tend not to make any further progress. The successful deployment of the machine-learning model takes into account:

   - Comprehensive understanding of the frontline users of model. The model deployment must be simple, intuitive, and user-friendly and packaged in simple-plain English tailored to how frontline users operate.
   - The deployed model should only have the necessary information for the frontline users to assist them in taking action or no action.
   - The frontline users should be able to process the insights from the deployed model at glance within 3 to 10 seconds.
   - Provide proper training and support for the frontline users during the model deployment.

Chapter 2

Basic Data Preparation

2.1 Setting Up the Retention Data

Before we can implement a retention analysis algorithm, it is necessary to set up the retention data and corresponding variables. In general, the structure of the data will have two distinct categories: the dependent variables and the independent variables.

The dependent variable represents the outcome we want to predict and to be consistent with machine learning terminology, we will refer to this variable as the response. In this case we want to predict whether a student is retained from an academic period to the next. For example, for a cohort of students beginning in the Fall of 2015 we might be interested if one year later they are retained. Thus, the value of the response variable, lets call it y, will be True if the student does *not* come back in Fall 2016 and False otherwise. In this case y is called a binary variable which can only take the value True or False. Sometimes we may want to represent True with the number 1 and False with the number 0.

The independent variables are those variables that influence the response variable y. Again, to be consistent with machine learning terminology, we will refer to this variable(s) as the feature(s). In retention analysis it is typical to use a combination of demographic and academic performance variables and metrics, lets call them X, where X refers to a tabular matrix where each row represents a student observation and each column represents a unique variable.

To illustrate this process, it is best to go through an example using data analysis tools available in the python ecosystem, namely Pandas:

2.1.1 Importing Data from a Comma Separated Values (CSV) File

In our example, we have a data file that contains retention data, including both the response variable and its features. The name of the file is 'retention_ data_ raw_sm .csv'. To read this file we import Pandas and then read the file:

```
1 import pandas as pd
2
3 df = pd.read_csv('retention_data_raw_sm.csv',index_col='EMPLID',)
4
5 print(df.shape)
6
7 df.head()
```

In the above code we use three common commands: read_ csv, print, and the method .head(). The command read_ csv reads the csv file and loads it into memory. Then we use print to print the shape (size) of the data table. And finally the .head() method is a command used to show the first few rows of a large dataset. In figure **??** we show the contents of the raw data. This particular

dataset has 4,506 rows (observations) and 8 columns (variables). Now the analyst is ready to pull the dependent variable an the set of independent variables need for the retention analysis.