Predictive Modeling of Student Dropout Using Machine Learning Approaches

RUTH NALUKWAGO[1], JANNIPHER NALUBEGA[2*], YASIR BUKENYA[3]

[1]DATA SCIENCE, Muteesa 1 Royal University, Mengo, Kampala, 34467, Uganda.
[2]DATA SCIENCE, Muteesa 1 Royal University, Masaka, Kampala, 07070, Uganda.

*Corresponding author(s). E-mail(s): janniphernalubega@gmail.com;
Contributing authors: nruth5455@gmail.com; alcantaradiaz28@gmail.com;

Abstract

Higher education student attrition isa complex problem affecting institutions and society. In this workwe set out to build a predictive model for identifying at-risk-dropout students based on 4,424 records (students) from an academic institute. The dataset contains 36 features that are grouped into demographic, academic, socioeconomic, and enrollment information of the students; here the target variableis a three-class label which indicates whether the student is a "Graduate," "Dropout" or "Enrolled".

In order to comprehend feature distributions and class imbalances, we conduct thorough exploratory data analysis (EDA), with graduates making up 49.9%, dropouts 32.1%, and enrolled students 17.9%. Various machine learning models were applied, including Logistic Regression, Random Forest, and XGBoost; the best performance was shown by the latter in predicting dropout likelihood. SMOTE tries to solve the problem of class imbalance

The final model provides actionable insights for early intervention that will enable the academic institution to provide needed support systems, improving student retention. This research has shown the potential of data-driven approaches in mitigating dropout rates and fostering academic success.

Keywords: Prediction on Dropouts of Students, Educational Data Mining, MachineLearning, XGBoost, Class Imbalance problem plus SMOTE

1 Introduction

The level of educational data in institutions of higher learning that is now very high has been a source of unforeseen opportunities for the study of students? behaviors and academic results. Access to detailed student records promotes empirical studies aimed at resolving critical educational challenges such as student performance prediction (1), dropout identification (2), and student clustering (3). Specifically, dropout prediction for students is the process of recognizing students in danger, depending on the historical patterns and their current academic performance, and, thus, it allows the teachers to make their interventions before students are disengaged.

Student characteristic (individuals) understanding academic performance determinants is an important sequence educational data set study. The student profile analysis feature importance is one of the major means to apply this (4). In this case, (5) suggests using multidimensional data sets, which include demographic, academic,

and socioeconomic information resources, to measure students' risk factors. Whereas, in (6)'s case, instead of behavioral centrality, the authors used academic performance metrics to maintain the student engagement structural patterns which were then applied to predictive academic outcomes.

In contrast to that, as varying student profiles usually exhibit some of the same traits in their academic journeys, investigators frequently make use of the similarity algorithms to replace the pattern information that is learned during model training in an optimal way. For instance, in (7), student classification is carried out with the help of educational features while the similarity between student profiles is measured by multiple academic factors, and finally, a hybrid classification criterion is applied in the prediction case.

But the above traditional methods of educational data mining rely on features engineered by humans and, therefore, are less effective in analyzing complex student behavior patterns. Inspired by the recent successes of deep learning in other domains, a few works extended approaches based on deep learning to educational data. (8).

Although standalone effective, such algorithms can hardly be combined with other factors to create a holistic view of student dropout applicable across institutions. We believe that a single solution can not fully solve this complex problem. To address these challenges, the Student Dropout Prediction (SDP) framework was developed. This generic machine learning approach combines various information sources, including demographic, academic performance, and socioeconomic characteristics, to model the interacting factors that cause students to discontinue studies. The strategy also uses more in-depth tactics like XGBoost (9) and tackles the imbalanced classes with SMOTE (10) for understanding the students' academic progression over time during model training.

Here's what we brought to the table with our work:

1. We dug into detailed student records, working with 36 different features that cover all sorts of angles.
2. We made sure to keep the tangled relationships between things like demographics, academics, and socioeconomic background front and center while training our model.
3. We tackled class imbalance and changing academic trends head-on, tying everything together in one clear framework.
4. We turned our findings into real advice schools can use to step in early and help students succeed.

2 Related Work

Research on predicting student dropouts has gone through a few big shifts. At first, people leaned on traditional stats and machine learning models—think Logistic Regression and Random Forests. These gave some clear insights, but you had to do a lot of manual work to set up the right features. Then came methods that looked for similarities and patterns in how students moved through their courses. These techniques picked up on trends in student data, but still needed a lot of hand-tuning. Lately, deep learning has entered the scene. These models can pick up complicated patterns on their own, but they run into trouble because education data is all over the place and not always plentiful.

In our work, we pull these different approaches together. We run a direct comparison of key algorithms like Logistic Regression, Random Forest, and XGBoost. Plus, we tackle real-world problems—like the fact that dropouts are pretty rare compared to students who stay—using tools like SMOTE to balance things out. The goal is to connect fresh methodological ideas with actual challenges faced in schools and universities.

3 Problem Definition and Preliminaries

3.1 Problem Definition

Here, I use bold capital letters like $\mathbf{X}$, vectors are denoted by boldface lowercase letters, e.g., $\mathbf{x}$, and scalars are indicated by lowercase letters, e.g., x. For ease of reference, the main mathematical notations deployed in this work are summarized in Table 1.

We start with a student dataset $\mathcal{D} = (\mathbf{X}, \mathbf{y})$. Here, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ holds n

Table 1: Here's a quick look at the main mathematical symbols used in this paper:.

Notation — What it means
n -Number of students in the dataset
m -Number of features for each student
$\mathbf{x}_i$ — Feature vector for the i^{th} student, lives in $\mathbb{R}^m$
$\widetilde{\mathbf{x}}_i$— Embedding vector for the i^{th} student, lives in $\mathbb{R}^d$ (with d smaller than m)
$\widehat{\mathbf{y}}_i$ — Predicted outcome for the i^{th} student, can be 0, 1, or 2
$\mathbf{X}$ — The full feature matrix, shape is $\mathbb{R}^{n \times m}$
$\mathbf{y}$ — Vector of the true outcomes, shape is $\mathbb{R}^n$
$\mathbf{W}$ — Weight matrix used in the classification model
$\mathcal{L}$— Loss function for training the model

student records, and each one packs m features—everything from demographics to academics and socioeconomic factors. Our labels, $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$, mark each student's outcome: $y_i \in \{\text{Dropout}, \text{Enrolled}, \text{Graduate}\}$.

Our job is to build a function $f : \mathbf{x_i} \longrightarrow \widehat{\mathbf{y}}_i$ that can look at a student's data and predict which category they land in. So, for the dropout prediction task, we want f to sort students into the right outcome—dropout, enrolled, or graduate—based on their records.

3.2 Feature Importance Analysis

Identifying the most influential features in a dataset is a cornerstone of educational data analysis. This process, known as feature importance analysis, measures how much each characteristic contributes to predicting student outcomes. The importance of a specific feature, denoted as I , can be calculated with the following formula:

$$I_j = \sum_{t=1}^{T} \sum_{\text{split } s \text{ on } j} \Delta\mathcal{L}_{s,t} \tag{1}$$

where T is the total number of trees in the ensemble model, and $\Delta\mathcal{L}_{s,t}$ represents the reduction in the loss function $\mathcal{L}$ when splitting on feature j at split point s in tree t.

3.3 Feature Similarity Measures

4 Proposed Methodology

Our proposed methodology follows a systematic approach to student dropout prediction:

4.1 Data Preprocessing

- Handling missing values through appropriate imputation techniques
- Encoding categorical variables using suitable encoding schemes
- Normalizing numerical features to ensure consistent scales
- Addressing class imbalance using SMOTE

4.2 Feature Engineering

We engineer features across multiple dimensions:

- Demographic features: age, gender, marital status, nationality
- Academic history: previous qualification grades, admission scores
- Socioeconomic factors: parental education, occupation, scholarship status
- In-program performance: course enrollments, grades, approval rates

4.3 Machine Learning Models

We implement and compare multiple machine learning approaches:

4.3.1 Logistic Regression

A baseline statistical model that provides interpretable coefficients.

4.3.2 Random Forest

An ensemble method that builds multiple decision trees for improved accuracy.

Table 2: Description of the feature similarity measures

Similarity Measure	**Description**	**Mathematical Formulation**
Academic Profile	Common academic features between students	$\sum_{k \in A} I(x_{ik} = x_{jk})$
Socioeconomic	Weighted socioeconomic similarity	$w^T s_{ij}$
Demographic	Demographic feature overlap ratio	$\frac{\sum_{k \in D} I(x_{ik}=x_{jk})}{\lvert D \rvert}$
Performance Trend	Academic performance correlation	Correlation coefficient
Risk Factor	Shared risk factors ratio	$\frac{2\sum_{k \in R} I(x_{ik}=x_{jk})}{\lvert R_i \rvert + \lvert R_j \rvert}$
Composite	Overall weighted similarity	$w^T sim_{ij}$

4.3.3 XGBoost

A gradient boosting framework known for its high performance and efficiency.

4.4 Model Evaluation

We employ comprehensive evaluation metrics:

- Accuracy, Precision, Recall, and F1-score
- Area Under ROC Curve (AUC-ROC)
- Cross-validation with stratified sampling
- Confusion matrix analysis

5 Experiments and Results

5.1 Dataset Description

Our experiments are based on the whole data set which contains 4,424 students with 36 features each.The distribution of the target variable is as follow::

- Graduate: 49.9% (2,209 students)
- Dropout: 32.1% (1,421 students)
- Enrolled: 17.9% (794 students)

5.2 Experimental Setup

Each of the experiments was done in Python and standard machine learning packages such as scikit-learn, XGBoost and imbalanced-learn. We have used 5-fold stratified cross validation to guarantee strong performance estimation

5.3 Results and Analysis

Our experimental results demonstrate that:

- XGBoost achieves the highest predictive performance among all tested models
- Addressing class imbalance through SMOTE significantly improves minority class recall
- Feature importance analysis reveals academic performance metrics as the strongest predictors
- The model provides actionable insights for targeted intervention strategies

Table 3: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.782	0.765	0.781	0.773
Random Forest	0.845	0.832	0.843	0.837
XGBoost	0.891	0.883	0.889	0.886

6 Conclusion

This work provided a complete machine learning model of student dropout prediction. Our solution adapts well to the problems of educational data such as imbalanced classes, heterogeneity of features and complicated temporal patterns. The outcomes of the experiment prove that our methodology will allow recognizing the at-risk students and providing the relevant interventions to decrease the retention rates.

Future work will explore:

- Inclusion of other time patterns and sequence modelling.
- Student segmentation using unsupervised learning.
- Creation of early warning real-time monitoring.
- Research into explainable AI methods of model explanations.

Acknowledgements

The study did not attract any grant, in either the public, commercial, or not-for-profit sector.

References

[1] Lu, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications.

[2] Bhagat, S., Cormode, G., & Muthukrishnan, S. (2011). Node classification in social networks. Social Network Data Analytics.

[3] Ebeid, I. A., & Fouda, M. M. (2022). Graph clustering algorithms: A comprehensive survey. Journal of Network and Computer Applications.

[4] Freeman, L. C. (1991). Centrality in social networks: Conceptual clarification. Social Networks.

[5] Huang, S., & Fang, N. (2014). Identifying student at-risk factors in educational data mining. Journal of Educational Technology & Society.

[6] Sserwadda, A., & Nakatumba-Nabende, J. (2021). Topology-based prediction of student academic performance. International Journal of Educational Technology in Higher Education.

[7] Rezaeipanah, A., & Ahmadi, G. (2019). Link prediction in social networks using topological features. Journal of Information Science.

[8] Monti, F., Bronstein, M., & Bresson, X. (2017). Geometric deep learning on graphs and manifolds using mixture model CNNs. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[9] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[10] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research.

[11] Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.