Engineering Educators Bringing the World Together
2025 ASEE Annual Conference & Exposition
Palais des congrès de Montréal, Montréal, QC • June 22–25, 2025 ASEE
Paper ID #47763

Engineering Student Early Dropout Prediction in Regional Universities Using Multimodal AI

Dr. Bin Chen, Purdue Univeristy Fort Wayne
Irah Modry-Caron, Purdue University Fort Wayne

Student Retention Forecast in Regional Universities

Introduction

The overall dropout rate of engineering students in the United States is approximately 50%. However, the dropout rate varies significantly across universities [1]. Prestigious national engineering schools often have retention rates over 90%. Regional universities and campuses have much higher student attrition rates. As a result, improving student retention has become a top priority in many regional universities.

Over the past decade, machine learning has received growing interest in college student retention prediction because more universities and institutions seek data-driven solutions to address this challenge. Researchers have applied various machine learning techniques in retention prediction, including decision trees, random forests, logistic regression, long short-term memory (LSTM) neural networks, as well as modern deep learning methods, such as AdaBoost, XGBoost, and lightGBM. According to a 2023 review paper by Attiya and Shams, the best out of the investigated algorithms in predicting student retention with performance rates is about 78% [2].

While many machine learning algorithms have shown promising results for retention prediction, most of these methods are designed for tabular data classification or regression. In tabular data classification, each student record is treated as an independent observation (row) with features represented as columns. The algorithms do not account for time dependencies between observations; they rely on a one-row, one-target training and validation approach. Timestamps, if included, are treated as one of the feature columns, with no specific consideration of the temporal relationships between observations. Predictions aim to assign a class label to each row without considering historical patterns [3-4]. Because tabular data classification typically delivers higher prediction accuracy, it has become the dominant approach for student retention prediction in recent years.

A more natural question in student retention prediction is whether students will remain enrolled in the following semester based on their historical data. This type of question falls outside the strengths of traditional tabular data classification methods. The data used for prediction are time series data where observations depend on each other, and past values influence future outcomes. The prediction task is to forecast the future value of the target variable based on part or all historical data, including students' enrollment history.

However, time series prediction methods are rarely applied to retention prediction because they are designed to predict numerical values instead of categorical classes. While time-series models have forecast capability, their prediction accuracies are often lower than those using tabular data prediction, making them less popular for this application.

In this study, we propose formulating the retention prediction as a classification task for sequential data. Traditionally, predicting a class from a time series sequence usually uses models like Hidden Markov Models, K-Nearest Models, and Support Vector Machines with sequence kernels. With the advancements in large language models, we introduce a new approach using the latest developments in multimodal AI to address this challenge. Historical student data is organized into time series sequences, and pre-trained transformer models are used for feature extraction and embedding. These embeddings are then passed to a classifier to predict a student's retention status at a future time point, such as next semester, depending on the prediction horizon [5-6].

This research study is approved by the university's Institutional Review Board (IRB). (The name is not revealed to comply with the double-blinded review guidelines.)

Dataset

A combination of data from the Office of Institutional Research's census and operational Banner data were preprocessed for anonymity and to comply with the requirements of IRB. Thirty-two thousand students registered for fall or spring semesters from Fall 2002 to Spring 2024 were included in this study. The dataset consists of five primary data sources: 1) high school information, 2) demographic information, 3) college and department program information, 4) academic information, and 5) financial information. This study selected a subset from each category, as shown in Table 1.

Table 1: Selected variables for student retention forecast.

Variable	Description
Student identifier	A unique number for student identification
Semester	Enrolled semester of students
Student classification	Freshman, sophomore, junior and senior
Enrollment status	Part-time or full-time students
First generation student	True or False
Gender	Male or Female
Ethnicity	White, Black or African American, Hispanic or Latino, Asian, Non-Resident Alien, American Indian or Alaskan Native, Native Hawaiian or Other Pacific Islander, Two or more races, Other.
Age	Student age
College code	Enrolled college

Department code	Enrolled department
Major code	Enrolled program
Highschool GPA	High School cumulative GPA
Cumulative GPA	Cumulative GPA up to a semester
Cumulative hours earned	Cumulative credit hours up to a semester
Number of withdraw grades, accumulated	Cumulative withdrew classes
Number of D grades, accumulated	Cumulative classes with grade D
Number of F grades, accumulated	Cumulative classes with grade F
Number of repeat courses, accumulated	Cumulative repeated classes
Number of GE courses, accumulated	Cumulative general education classes
Lower division courses	Courses that are typically taken by freshmen and sophomores. 100–200 level courses.
Upper division courses	Courses that are typically taken by juniors and seniors. 300–400 level courses.
Distance from campus	Distance from home to campus for students who commute.
isDropout, isStopout	Status of dropout or stopout
Semester enrollment count	Cumulative semesters
Degree complete	True or False

One critical piece of student information not included in this draft is financial status. Due to time constraints and the complexity of financial data, the authors could not incorporate data such as loans, financial aid from local, state, and federal sources, scholarships, awards, family support, or employer contributions.

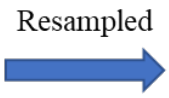
Preprocessing

The data was organized in the panel data format. Each item is a table of timeseries. The variables "dropout," "isStopout," "Semester enrollment count," and "Degree Complete" listed in Table 1 were used to derive a new variable, "Enroll Status." This new variable represents a student's enrollment status for each semester from the beginning to the end, with the exit status of either "dropout" or "graduate." In this study, "dropout," "stop out," and "graduate" are grouped as "not_enrolled" because the goal of the project is to predict if a student will enroll next semester. This simplifies the problem into binary classification tasks. The texts in the "Semester" column were converted to timestamps, and the "Student ID" column is treated as item ID for panel data. To ensure a constant frequency of timestamps, the data were resampled to include all spring and fall semesters from Fall 2002 to Spring 2024 (Figure 1).

Original

Student ID	Semester	Status
11015	Fall 2002	enroll
11015	Spring 2003	enroll
11015	Fall 2003	enroll
11015	Spring 2004	enroll
11015	Spring 2005	enroll
11015	Fall 2005	enroll
11015	Spring 2006	enroll
11015	Fall 2006	enroll
11015	Spring 2007	enroll
11015	Fall 2007	enroll
11015	Spring 2008	graduate

Resampled



Resampled

Student ID	Semester	Status
11015	Fall 2002	enroll
11015	Spring 2003	enroll
11015	Fall 2003	enroll
11015	Spring 2004	enroll
11015	Fall 2024	stopout
11015	Spring 2005	enroll
11015	Fall 2005	enroll
11015	Spring 2006	enroll
11015	Fall 2006	enroll
11015	Spring 2007	enroll
11015	Fall 2007	enroll
11015	Spring 2008	graduate

Figure 1: A sample student enrollment record. Data resampled to include stopout semesters.

Retention Forecast

Figure 2 shows the time series sequences as input to the deep learning network. These sequences were embedded into a fixed length for easy batching during training to ensure compatibility with standard deep-learning training strategies. Once the data are correctly fed into the transformer-based model, techniques commonly used in state-of-the-art deep learning can be applied to student retention prediction. This method makes the application more generalized, more robust, and less overfitting. Figure 3 shows the machine learning pipeline.

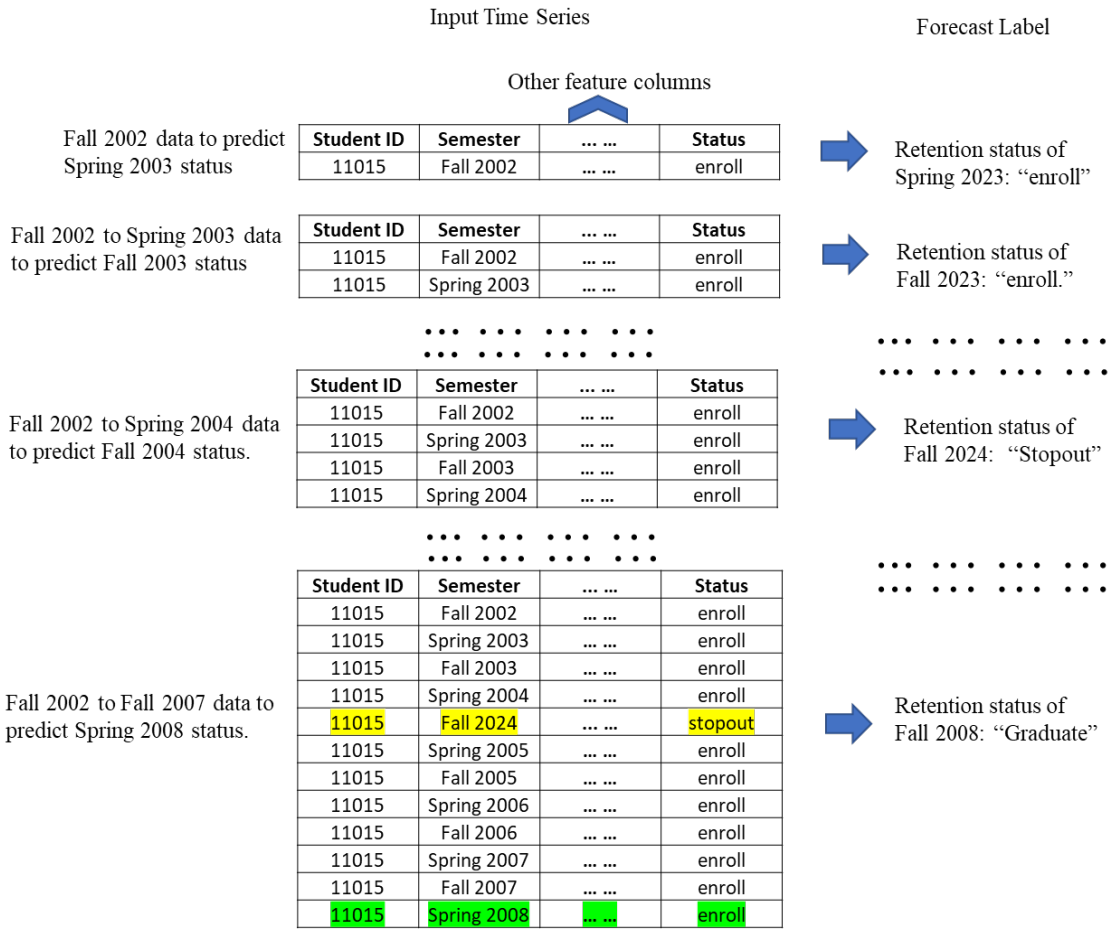


Figure 2: Time series forecast with 1 horizon, predicting following semester's retention status of a student from all historical data of that student. Change prediction horizon for longer time forecast.

The results in Table 2 show that the overall accuracy of predicting whether a student will enroll in the next semester is approximately 82.88%. This accuracy is comparable to the performance achieved using tabular data prediction methods. However, unlike tabular data prediction, this

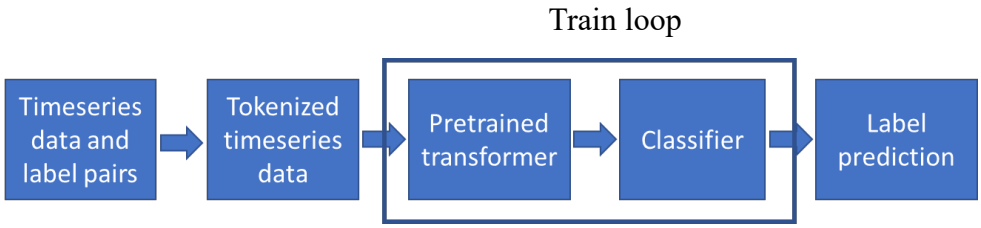


Figure 3: Pipeline of sequence prediction. The pretrained transformer was finetuned and the classifier was trained for student retention prediction using the standard deep learning training, validation and testing procedures.

approach offers forecasting capabilities, enabling predictions for future time points, even progressively as new data becomes available.

Table 2: Student retention prediction results

--------- Evaluation Metrics -----------
Accuracy = 82.88 %

	precision	recall	f1-score
not_enroll	0.82	0.45	0.58
enroll	0.83	0.96	0.89

Limitations

The dataset used in this study does not contain all relevant variables of interest to predict student retention. Understanding these limitations is crucial for interpreting the findings accurately and identifying areas for future research and improvement. In particular, the dataset does not include financial aid information, which restricts the ability of financial aid staff to utilize the model's output for making informed financial aid allocation decisions. The dataset contains limited student demographic information. Expanding the range of student characteristics included in the dataset could enable the creation of more targeted intervention strategies. It is well-documented that factors such as being an honors student, a student-athlete, or participating in student government significantly contribute to student retention. Incorporating these student behavioral characteristics into the model would enhance the overall accuracy of retention forecasts.

Intervention Strategies

Retention forecasts could inform financial aid allocation decisions if financial aid data were included in the study. Specifically, the university could create financial aid packages that incentivize students to make progress toward earning a bachelor's degree. Additionally, exploring financial aid optimization tools and technologies could support this effort.

Accumulating many W (withdrawal) and F (fail) grades is a significant risk factor for dropping out. Since many new students take general education courses during their first semester, it is concerning if they have a significant number of W and F grades. To address this issue, the university should consider reconfiguring academic support systems to focus on high DFW courses and training faculty to identify and refer students who need extra academic support during their first year.

Research indicates that the distance students travel to campus affects their graduation chances. Providing on-campus housing is crucial for promoting retention and timely graduation. Therefore, the university should consider expanding its on-campus housing facilities and integrating housing with academic, curricular, and social activities to keep students engaged.

Beyond housing, fostering community and belonging is essential for student retention. The university could develop programs encouraging student participation in clubs, organizations, and

campus events. Mentorship programs pairing new students with upperclassmen or faculty advisors could provide additional support and guidance.

Conclusions

This proposed framework takes advantage of the latest progress in AI. Pretrained transformers are highly versatile, delivering excellent performance in various tasks from chatbot agents to image generation. A pre-trained network also significantly reduces the required data, enabling robust performance even with limited samples. Additionally, this pipeline is highly scalable and adaptable. It can incorporate various types of student information, including structured data from administrative systems (e.g., Banner), unstructured data such as open-ended survey responses, and even images or drawings, improving retention predictions' accuracy and reliability.

References

1. Sean Kim, Eliot Yoo, Samuel Kim. Why Do Students Drop Out? University Dropout Prediction and Associated Factor Analysis Using Machine Learning Techniques. 2023. https://arxiv.org/abs/2310.10987
2. W. M. Attiya and M. B. Shams, "Predicting Student Retention in Higher Education Using Data Mining Techniques: A Literature Review," *2023 International Conference On Cyber Management And Engineering (CyMaEn)*, Bangkok, Thailand, 2023, pp. 171-177.
3. N. S. Sani, A. F. M. Nafuri, Z. A. Othman, M. Z. A. Nazri, and K. Nadiyah Mohamad, “Drop-Out Prediction in Higher Education Among B40 Students,” International Journal of Advanced Computer Science and Applications, vol. 11, no. 11, pp. 550–559, 2020
4. D. E. Moreira da Silva, E. J. Solteiro Pires, A. Reis, P. B. de Moura Oliveira, and J. Barroso, “Forecasting Students Dropout: A UTAD University Study,” Future Internet, vol. 14, no. 3, Mar. 2022, doi: 10.3390/fi14030076.
5. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).
6. Transformers. https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/