AI-Driven Strategies for Reducing Student Withdrawal

A Study of EMU Student Stopout

Yan Zhao*, Amy Otteson *

*Eastern Michigan University, USA

Abstract—Not everyone who enrolls in college will leave with a certificate or degree, but the number of people who drop out or take a break is much higher than experts previously believed. In December 2013, there were 29 million people with some college education but no degree. That number jumped to 36 million by December of 2018, according to a new report from the National Student Clearinghouse Research Center[1].

It is imperative to understand the underlying factors contributing to student withdrawal and to assist decision-makers to identify effective strategies to prevent it. By analyzing the characteristics and educational pathways of the stopout student population, our aim is to provide actionable insights that can benefit institutions facing similar challenges.

Eastern Michigan University (EMU) faces significant challenges in student retention, with approximately 55% of its undergraduate students not completing their degrees within six years. As an institution committed to student success, EMU conducted a comprehensive study of student withdrawals to understand the influencing factors. And the paper revealed a high correlation between certain factors and withdrawals, even in the early stages of university attendance. Based on these findings, we developed a predictive model that employs artificial intelligence techniques to assess the potential risk that students abandon their studies. These models enable universities to implement early intervention strategies, support at-risk students, and improve overall higher education success.

I. INTRODUCTION

Student withdrawal from college is a critical issue that significantly impacts higher education institutions across the United States. Despite growing enrollment rates, a considerable number of students leave college without completing their degrees. This trend has raised concerns among educational policy makers and institutions.

Addressing the issue of student withdrawal is essential for improving graduation rates and ensuring student success. By identifying the underlying factors that contribute to student withdrawal and developing strategies to address these factors, institutions can better support their students. This study aims to provide a comprehensive analysis of the stop-out student population at Eastern Michigan University (EMU) to understand the characteristics and educational pathways that lead to withdrawal.

Specifically, this study focuses on two primary research questions concerning new First-Time In Any College (FTIAC) students.

- **Q1: Which students are most likely to withdraw from school, and what common characteristics do they share?**

- **Q2: Can these common characteristics be used to predict the risk of student withdrawal?**

For Q1, the analysis includes various factors such as academic performance, demographics, admission criteria, financial status, and other relevant characteristics to understand the predictors of student withdrawal. In order to quantify the stopout group, retention rates are used as metrics. Retention rates measure the percentage of first-time undergraduate students who return to school in subsequent years. For Q2, the study examines whether the common characteristics identified in Q1 can be utilized to predict the risk of student withdrawal. By leveraging machine learning techniques, predictive models are developed to identify students at high risk of stopping out. These models use identified predictors, such as academic performance, demographics, institutional factors, engagement levels, and financial aid status, to assess the likelihood of student withdrawal. The goal is to provide actionable insights that enable EMU to implement targeted interventions and support strategies to improve student retention and success.

The data drawn from EMU's enrollment and degree records between Fall 2013 and Fall 2017 provide a foundation for exploring these questions. The insights gained from this analysis are expected to inform strategies that can improve student retention rates, not only at EMU but also at other institutions facing similar challenges.

II. RELATED WORK

The issue of college student withdrawal has significant implications for both students and educational institutions. Extensive research has been conducted to understand the factors that contribute to student attrition and identify effective strategies to improve retention rates. This section reviews the relevant literature and previous studies on various dimensions of student withdrawal.

Previous research consistently shows that demographic characteristics and academic performance are critical predictors of student withdrawal. Tinto [2] emphasized the importance of academic integration, suggesting that students who perform well academically and feel integrated into the academic community are less likely to withdraw. Herzog [3] found that GPA and credit hours completed are significant predictors of student persistence.

Astin's theory of student involvement [4] and Kuh et al.'s emphasis on high-impact educational practices [5] highlight

1

the importance of student engagement in retention. Astin's theory of student involvement [4] and Kuh et al.'s emphasis on high-impact educational practices [5] highlight the importance of student engagement in retention.

Our study extends this by integrating correlation analysis to offer a holistic understanding of student withdrawal, combining quantitative rigor with qualitative depth. And instead of focusing on one single factor, our research utilizes a comprehensive dataset from Eastern Michigan University (EMU) to explore the interplay between academic performance, demographic characteristics, and admission criteria.

Stratton, O'Toole, and Wetzel [6] found that students are most vulnerable to withdrawal during their first year of college. Pascarella and Terenzini [7] noted that the first year is essential to establish academic habits and social connections.

Unlike previous studies, our research incorporates a predictive model to analyze the risk of student withdrawal at the early stage of university attendance, providing valuable information on potential at-risk students.

These comprehensive approaches offer specific insights that can directly inform policy and practice at EMU and have broader applications for similar institutions.

Zhao [8] utilized Markov Chain models for enrollment prediction, demonstrating the potential of advanced analytical techniques in educational research. While this research is useful for identifying early warning trends of withdrawal, it does not provide the capability to pinpoint individual at-risk students.

III. DATA COLLECTION

This study aims to analyze the factors that contribute to the withdrawal of students from Eastern Michigan University (EMU) and to identify common characteristics among students who are more likely to leave without completing their degree.

We use retention rates as metrics to determine whether students stay or leave the school, rather than relying on degree completion rates. Retention rates measure the percentage of first-time undergraduate students who return to school in subsequent years. Avg FA #2, Avg FA #3, Avg FA #4, Avg FA #5, and Avg FA #6 represent the average proportions of freshmen entering in Fall 2013 through Fall 2017 at EMU who returned in the second, third, fourth, fifth and sixth fall, respectively. We chose not to use final degree completion rates directly because our goal is to identify at-risk students as early as possible. By doing so, the university can implement timely interventions to support these students and improve their chances of success.

Meanwhile, we collect data regarding academic performance, specifically term GPA and term earned Student Credit Hours (SCH), admission criteria like ACT score and decision GPA and demographic information.

Data were derived from EMU's enrollment and degree records between Fall 2013 and Fall 2017.

- **Retention Rates**: Percentage of first-time undergraduate students returning in subsequent years.
- **Term GPA**: Average GPA of students categorized into five ranges: >3.5, 3.5-3.0, 3.0-2.0, 2.0-1.0, and <1.0.

- **Term Earned SCH**: Number of credit hours completed in the first fall term, categorized into four ranges: >12, 9-12, 6-9, and <3.
- **ACT Scores**: Composite scores categorized into four ranges: >30, 24-29, 18-23, and 12-17.
- **Decision GPA**: High school GPA used for college admissions.
- **Avg Fall #2** : the average proportions of freshmen entering in Fall 2013 through Fall 2017 at EMU who returned in the second year.

IV. CORRELAITON EXPERIMENTS

This section presents our investigation for the first research questions RQ1. we will first introduce the study method and then describe the results of our experiment.

RQ1: Which students are most likely to withdraw from school, and what common characteristics do they share?

A. Pearson Correlation Coefficient Calculation

The Pearson correlation coefficient, denoted as r, quantifies the linear relationship between variables:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Calculating the Pearson correlation coefficient provides several key benefits:

- **Quantifies Relationships**: It gives a precise measure of the strength and direction of the relationship between term GPA and retention rates.
- **Statistical Validation**: Provides formal statistical validation of observed trends.
- **Standardized Comparison**: Allows for consistent comparison across different studies and datasets.
- **Confidence Intervals**: Offers insights into the reliability and precision of the correlation estimate.
- **Objective Assessment**: Minimizes bias by providing an objective, mathematical assessment of the relationship.

B. Academic Standing (GPA and SCH)

Academic standing is a significant predictor of whether students are likely to stop out. We measured academic standing using two categories: term GPA and term-earned SCH in the first fall.

1) Term GPA and Retention Rate Data: Figure 1 shows the retention rates in different term GPA ranges over the first four years:

- 89%, 80%, and 74% for students with a term GPA > 3.5
- 83%, 70%, and 63% for students with a term GPA between 3.5 and 3.0
- 73%, 54%, and 45% for students with a term GPA between 3.0 and 2.0
- 43%, 26%, and 22% for students with a term GPA between 2.0 and 1.0
- 13%, 6%, and 6% for students with a term GPA below 1.0

Figure 2 shows the retention rates in the different term-earned SCH groups over the next four years are as follows:

Fig. 1: Retention Rates over Six Years by Term GPA

- 76%, 65%, and 58% for students who completed more than 12 credit hours in the first fall
- 65%, 49%, and 42% for students with term-earned SCH between 9 and 12
- 41%, 30%, and 26% for students with term-earned SCH between 6 and 9
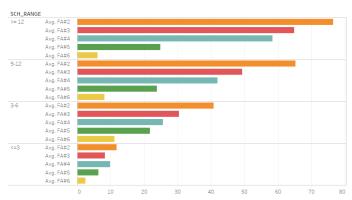- 12%, 8%, and 9% for students with term-earned SCH below 3

Fig. 2: Retention Rates over Six Years by SCH

2) Correlation Results: To calculate the correlation, we use the following data points:

- For Term GPA and Retention Rates:
  - Term GPA: [4.0, 3.25, 2.5, 1.5, 0.5] (representing GPA ranges: >3.5, 3.5-3.0, 3.0-2.0, 2.0-1.0, <1.0)
  - Second FAll Retention Rates: [0.89, 0.83, 0.73, 0.43, 0.13]
- For Term Earned SCH and Retention Rates:
  - Term Earned SCH: [13.5, 10.5, 7.5, 1.5] (representing SCH ranges: >12, 9-12, 6-9, <3)
  - Sencond Fall Retention Rates: [0.76, 0.65, 0.41, 0.12]

Using these data points, we calculate the Pearson correlation coefficients.

Correlation between Term GPA and Retention Rates:
Pearson correlation coefficient: $r = 0.993$
95% confidence interval: $(0.979, 1.008)$

Correlation between Term Earned SCH and Retention Rates: **Pearson correlation coefficient:** $r = 0.993$
95% confidence interval: $(0.974, 1.012)$

These high correlation values indicate a very strong positive relationship between both term GPA and retention rates, and term earned SCH and retention rates. This suggests that higher term GPAs and higher term earned SCH are strongly associated with higher retention rates at EMU. The rates at which students returned to EMU increased with the quality and amount of academic progress made during the first fall.

3) Key Insights: Several key insights can be gained regarding the relationship between academic standing and student retention rates over the first four years:

- **Higher GPA and SCH Correlates with Higher Retention Rates**: Students with higher term GPAs and SCH in their first fall semester are significantly more likely to stay enrolled in subsequent years.
- **Declining Retention with Lower GPA and SCH**: There is a clear stepwise decline in retention rates as the term GPA and SCH decrease.
- **Severe Drop in Retention for Lowest GPA Bracket**: Students with term GPAs between 2.0 and 1.0 have markedly lower retention rates, with only 43% returning in the second year. The situation is most critical for students with a term GPA below 1.0, where retention rates are dramatically low at 13% in the second year. Students with SCH below 6 show the same trend.

These insights suggest that academic performance in the first term is a strong predictor of student retention. Targeted interventions could be especially beneficial for students with GPAs below 2.0 or SCH below 6 to help improve their academic standing and increase their likelihood of persisting in their studies.

Understanding these patterns can help universities allocate resources more effectively. Institutions can prioritize academic support services for students identified as high-risk based on their first-term GPA and SCH.

In summary, the data in Figures 1 and 2 highlight the critical importance of early academic success in predicting student retention, underscoring the need for early intervention strategies to support at-risk students and improve overall retention rates.

The difference in retention rates in the fifth and sixth falls is not as significant as in previous years because students with higher term GPAs and more term earned SCH have higher completion rates within four years.

C. Admission Criteria (Test Scores and Decision GPA)

Test scores and decision GPA (high school GPA) are two important college-admissions criteria.
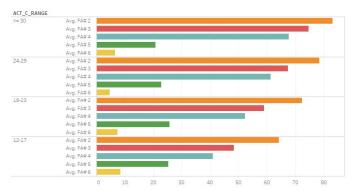
Fig. 3: Retention Rates over Six Years of ACT Score Group

As shown in Figure 3. Specifically, 83% students with an ACT Composite score greater than 30 re-enrolled in the second fall, while 78% students with ACT Composite score between 24 and 29 were retained. Of the students whose composite ACT score was between 18 and 23, 73% of the students reenrolled. Conversely, only 64% students with ACT Composite score between 12 and 17 re-enrolled. They also keep the same trend in the next four fall terms. The retention rates for the third fall were 75%, 67%, 59%, and 48% and 68%, 61%, 52%, and 49% for the fourth fall for the four groups of undergraduate students, respectively.

> *Correlation between ACT score and Retention Rates:*
> **Pearson correlation coefficient:** $r = 0.992$
> **P-value:** p=0.0075

The correlation coefficient r = 0.992 indicates a very strong positive correlation between the ACT composite scores and retention rates for the second fall term. The p-value p = 0.0075 is well below the common significance threshold of 0.05, indicating that this correlation is statistically significant.
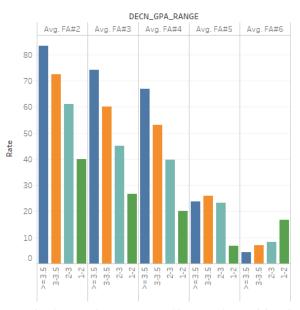

Fig. 4: Retention Rates over Six Years by Decision GPA

Figure 4 below shows the distribution of retention rates in different decision GPA groups. Specifically, we see that the retention rates in the first four years are 83%, 74%, and 67% for students with a decision GPA greater than 3.5, decreasing to 72%, 60%, and 53% for students with decision GPA between 3.0 and 3.5. For students with a decision GPA between 2.0 and 3.0, 61%, 45%, and 40% were retained. There was a large drop in retention rates for students with a decision GPA of 1.0 to 2.0: 40%, 27% and 20%. From the fifth fall, the retention rates are not consistent with previous years because students with higher ACT Composite scores or better decision GPA are more likely to earn their degree in four years.

> *Correlation between Decision GPA and Retention Rates:* **Pearson correlation coefficient:** $r = 0.995$
> **P-value:** p=0.0049

This very high Pearson correlation coefficient 0.995 and low p-value (< 0.05) indicate that the correlation between decision GPA and retention rates in the first year is statistically significant, strongly supporting the hypothesis that higher decision GPAs are associated with higher retention rates.

Based on these insights, institutions can design targeted interventions to support students with lower ACT scores, potentially improving their retention rates. Admission policies can take these findings into account to ensure that students who may need additional support are identified early and provided with the necessary resources to succeed and stay enrolled.
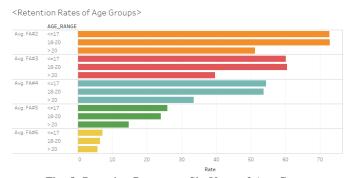
D. Demographic Information

Fig. 5: Retention Rates over Six Years of Age Group

The vast majority of FTIAC students (84%) are between the age of 18 and 20, with just 15 percent below 17 years of age. Only one percent are older than 20 years. A closer look at the retention rates of different age groups shows that there is no difference between students below age 17 and students age 18-20, with their retention rates (73% vs. 73% for the second fall, 60% vs. 60% for the third fall, and 54% vs.54% for the fourth fall). Students under 20 years old achieve higher retention rates than students over 20, as seen below in Figure 5. The retention rates are 72% vs. 51% (second fall), 60% vs. 39% (third fall) and 54% vs. 33% (fourth fall), respectively.

The negative correlation coefficients of -0.866 suggest a strong inverse relationship between age groups and retention rates, indicating that younger students (below 20) tend to have higher retention rates compared to older students (over 20). However, the p-values of 0.333 indicate that these correlations are not statistically significant at the conventional alpha level of 0.05. This means that we cannot confidently conclude that there is a significant correlation between age groups and retention rates based on these data. The evidence supporting the influence of age as a factor is sufficient.

We also analyzed the variables of gender and race. Similarly to age, their correlations with student withdrawal were not statistically significant.

E. Finacial Status

The dataset contains the following key variables:
LOW INCOME: Indicator of whether a student is from a low-income family (Y/N).
FA#2 to FA#6: Retention rates at different follow-up fall semester. The data is summarized in Table I

LOW INCOME	FA#2	FA#3	FA#4	FA#5	FA#6
Y	73.93	61.47	54.13	32.73	14.33
N	71.71	59.33	53.19	27.85	7.86

TABLE I: Retension rates by Low Income Status

Contrary to some expectations and existing research, our dataset indicates that low-income students exhibit slightly higher retention rates at various follow-up points compared to their non-low-income peers. A plausible explanation for this trend could be targeted support programs for low-income students at Eastern Michigan University (EMU). For students who receive financial aid, they are more inclined to maintain their aid by continuing their studies.

This factor is heavily influenced by the specific policies and criteria of financial aid, which can vary significantly from one institution to another. As a result, it is difficult to generalize this finding to other schools. In the subsequent predictive models, we will exclude this factor due to its variability and the challenges in applying it consistently across different educational contexts.

V. PREDICTIVE MODEL

A. Model Development

The objective of this section is to develop predictive models for student retention, especially focusing on improving performance for students who are likely to drop out. This task is challenging due to the imbalanced nature of the dataset, where the number of retained students is significantly greater compared to the number of non-retained students. Imbalanced datasets can lead to biased models that perform well on the majority class but poorly on the minority class. To address this, we employed various techniques including advanced models like XGBoost. XGBoost is a powerful and efficient implementation of gradient boosting that often performs well on classification tasks, especially with imbalanced datasets.

1) Data Preparation: The dataset includes 13995 FTIAC student records, and each record has the following feature:

- **Term GPA**: Term GPA in the first fall term.
- **Term Earned SCH**: Number of credit hours completed in the first fall term.
- **ACT Scores**: ACT composite scores.
- **Decision GPA**: High school GPA used for college admissions.

The target variable is the second fall retention, with a value of 1 indicating that a student is retained, and a value of 0 indicating that a student has stopped out. Rows with missing values in the selected features and target were removed to ensure clean data. The data was split into training (80%) and testing (20%) sets to evaluate model performance.

2) Model Training with XGBoost: XGBoost is an advanced gradient boosting algorithm known for its efficiency and high performance. It is particularly effective in handling imbalanced datasets through its ability to use various parameter tuning options to handle class imbalance. Furthermore, the SMOTE technique was applied to ensure that the minority class (retention = 1) was adequately represented.

B. Results

The model has been trained and evaluated. The performance of the model is summarized in Table II.

Class	Precision	Recall	F1-Score	Support
0	0.835	0.868	0.851	1716
1	0.588	0.523	0.553	616
Accuracy			0.777	2332
Macro Avg	0.711	0.696	0.702	2332
Weighted Avg	0.770	0.777	0.773	2332

TABLE II: Classification Report for the Trained Model

C. Model Performance

The overall accuracy of the model is 77.7%, which reflects a balanced performance across both classes. The results indicate that the model performs well in identifying the students who are retained (class 0), with high precision and recall values. For students who are not retained (class 1), the precision and recall are slightly lower, suggesting room for improvement in accurately predicting this class.

The macro and weighted averages provide additional insight into the model's balanced performance across different metrics.

The macro average provides an unweighted mean of the precision, recall, and F1-score for each class. This metric treats all classes equally, regardless of their support (the number of instances in each class). It is useful for understanding the model's performance on a per-class basis, without being influenced by class imbalance.

The weighted average takes into account the support of each class, providing a mean that is weighted by the number of instances in each class. This metric is particularly useful in the presence of class imbalance as it reflects the

model's performance more accurately in terms of the overall population.

The choice of model and threshold significantly impacts the recall and precision rates for students who are not retained (class 1). By selecting different thresholds, we can adjust the balance between recall and precision to suit various needs:

- **High Precision Model**: Useful for targeted interventions where resources are limited and must be used efficiently.
- **High Recall Model**: Suitable for broad preventive measures aiming to cover as many at-risk students as possible.

D. Implementation

We used data from Eastern Michigan University (EMU) to develop our predictive model, but other universities can adapt and implement this model to identify students at high risk of withdrawal. By leveraging this model, institutions can deploy early intervention strategies, offering targeted support services such as academic advising, tutoring, financial aid counseling, and engagement programs, ultimately aiming to improve student retention rates.

VI. CONCLUSION

This study demonstrates the feasibility of using data-driven approaches to understand student withdrawal. By leveraging academic performance metrics, demographic information, and admission criteria, EMU can develop effective strategies to identify at-risk students in early stage and enhance overall student success.

The analysis of the characteristics and educational pathways of the stopout student population at Eastern Michigan University (EMU) has yielded valuable insights. The primary findings of this study are as follows.

A. *Academic Performance as a Predictor of Retention*

There is a strong positive correlation between academic performance metrics (such as term GPA and term earned SCH) and student retention rates. Students with higher term GPAs and earned SCH in their first fall semester are significantly more likely to stay enrolled in subsequent years. This finding underscores the critical importance of early academic success in predicting student retention.

B. *Impact of Admission Criteria*

Higher ACT scores and decision GPAs are associated with higher retention rates. This correlation indicates that students who perform well in standardized tests and high school are more likely to persist in their studies at EMU.

C. *Demographic Factors*

Age, gender and race did not show statistically significant correlations with student withdrawal. These results suggest that targeted interventions may not be necessary for older students or a specific gender to help improve their retention rates.

D. *Predictive Modeling*

The XGBoost model developed in this study effectively predicts the likelihood of student withdrawal based on academic performance metrics and admission criteria. The high accuracy, precision and recall values of the model indicate its robustness in identifying at-risk students.

VII. IMPLICATIONS FOR EMU AND SIMILAR INSTITUTIONS

The insights gained from this study can inform strategies to improve student retention rates at EMU and other institutions facing similar challenges. By identifying the key predictors of student withdrawal, universities can allocate resources more effectively and implement early intervention strategies to support at-risk students.

These strategies may include the following:

- Providing academic support services such as tutoring and academic advising to students with low-term GPA and earned SCH.
- Offering financial aid counseling to ensure that students receive adequate financial support.
- Designing engagement programs to help integrate students into the academic community, particularly during their first year of college.

VIII. FUTURE RESEARCH DIRECTIONS

While this study provides a comprehensive analysis of student withdrawal at EMU, future research could explore additional factors that may influence student retention, such as finance, social integration, and external commitments (e.g., employment). Longitudinal studies could also examine the long-term impact of early interventions on student success.

In conclusion, this study highlights the importance of data-driven approaches in understanding student withdrawal. By leveraging predictive modeling and explainable AI techniques, EMU and other higher education institutions can develop effective strategies to support at-risk students and improve overall student success. The findings from this research provide a solid foundation for ongoing efforts to enhance student retention and ensure that more students achieve their academic goals.

REFERENCES

[1] N. S. C. R. Center, "Some college, no degree: A 2019 snapshot for the nation and 50 states," 2019. [Online]. Available: https://nscresearchcenter.org/wp-content/uploads/SCND_Report_2019.pdf

[2] V. Tinto, *Leaving College: Rethinking the Causes and Cures of Student Attrition*. University of Chicago Press, 1993.

[3] S. Herzog, "Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen," *Research in Higher Education*, vol. 46, no. 8, pp. 883–928, 2005.

[4] A. W. Astin, "Student involvement: A developmental theory for higher education," *Journal of college student personnel*, vol. 25, no. 4, pp. 297–308, 1984.

[5] G. D. Kuh, *High-impact educational practices: What they are, who has access to them, and why they matter*. Association of American Colleges and Universities, 2008.

[6] L. S. Stratton, D. M. O'Toole, and J. N. Wetzel, "Dropout, stopout, and transfer: The impact of attendance patterns on college student success," *Research in Higher Education*, vol. 48, no. 4, pp. 433–465, 2007.

[7] E. T. Pascarella and P. T. Terenzini, *How College Affects Students: A Third Decade of Research*.   Jossey-Bass, 2005.

[8] Y. Zhao and A. Otteson, "A practice in enrollment prediction with markov chain models," *arXiv preprint arXiv:2405.14007*, 2024.