International Journal of Computer Applications Technology and Research
Volume 13–Issue 08, 54 – 61, 2024, ISSN:-2319–8656
DOI:10.7753/IJCATR1308.1006

Role of Advanced Data Analytics in Higher Education: Using Machine Learning Models to Predict Student Success

[a] Harold Tobias Adu-Twum, Department of Mathematics and Statistics, Youngstown State University, USA,; [b] Emmanuel Adu Sarfo, Department of Mathematics and Statistics, Youngstown State University USA, [c] Evans Nartey, [d] Adesola Adetunji, Department of Chemical Engineering, University of North Dakota; [e] Adebowale Olufemi Ayannusi, Department of Computer Science, Ogun State Institute of Technology, [f] Thomas Andrew Walugembe, Maharishi International University, Fairfield Iowa

Corresponding Author Contact: Harold Tobias Adu-Twum Department of Mathematics and Statistics, Youngstown State University, Youngstown, 1 Tressel Way, OH, USA, 44555

ABSTRACT

This research article explores the pressing issue of college student dropout, employing a suite of predictive modeling techniques to forecast students' likelihood of discontinuing their studies. In an era where higher education plays a crucial role in individual career prospects and societal progress, understanding and mitigating dropout rates is of paramount importance. We leverage a comprehensive dataset, incorporating demographic, socioeconomic, academic, and financial factors, to train and test predictive models: Logistic Regression, Random Forest, Decision Tree Classifier, Support Vector Machine (SVM), and Gradient Boosting.

Our analysis reveals that the gradient boosting model outperforms its counterparts in predicting student dropout, achieving the highest precision, recall, and F1 score among the evaluated models. Gradient boosting model correctly identified 94.4% of the student as dropouts. This superiority underscores the gradient boosting's robustness in handling the complex, multidimensional nature of factors influencing college retention. Moreover, it was found that circular units approved in the first semester is the most important factor in determining student success in terms of graduating or dropping out. This is then followed by tuition fees up to date. The previous qualification of students has the lowest predictive power indicating that current circumstances of students – academic courses, and finances contributes to students' success more. The findings underscore the potential of machine learning in crafting targeted interventions to support at-risk students, thereby enhancing retention rates and contributing to the broader objectives of educational equity and success.

Keywords: Predicting Student Dropout, Binary Classification, Machine Learning in Higher Education, Boosting Methods, Advance Data in Higher Education, Student Retention

1.0 INTRODUCTION

College students face numerous challenges, often leading to a pivotal decision: whether to persist and graduate or to disengage and drop out. Understanding the factors that influence this decision is critical for educational institutions aiming to improve student retention rates and support academic success. This research article introduces a comprehensive predictive modeling approach to identify the key determinants of college student retention. By leveraging a rich dataset encompassing demographic, socioeconomic, academic, and psychosocial variables, our study employs advanced statistical and machine learning techniques to forecast the likelihood of a student either dropping out or graduating from college.

The significance of addressing college dropout rates extends beyond individual outcomes, impacting societal welfare, workforce development, and the broader educational ecosystem. Previous research has highlighted various predictors of college dropout, including academic preparedness, financial constraints, and social integration, yet the complexity of student retention demands a multifaceted analysis that considers the interplay between these factors. Our study responds to this need by integrating diverse predictors into a holistic model, offering insights into the nuanced dynamics of student persistence.

By identifying students at high risk of dropping out, our predictive model aims to enable targeted interventions by educational institutions. These interventions can be customized to address the specific needs and challenges faced by at-risk students, thereby enhancing their potential for academic success. Furthermore, our research contributes to the theoretical understanding of student retention, providing a foundation for future studies in this domain.

In summary, this article presents an effort to predict college student dropout and graduation outcomes through a multidisciplinary approach. The findings not only have practical implications for improving retention strategies but also enrich the academic discourse on navigating the complexities of higher education.

1.1 Problem Statement and Research Objective

The dropout rate of college students not only affects the students' future career prospects and personal growth but also has broader societal implications, including economic repercussions and the underutilization of human capital. Despite numerous studies and interventions aimed at understanding and reducing dropout rates, the persistence of this issue highlights a complex interplay of factors that are not fully understood or addressed by current models and strategies. Therefore, there is a critical need for more sophisticated predictive models that can accurately identify at-risk students based on a comprehensive set of predictors, enabling timely and targeted interventions.

The primary objectives of this research are as follows:

1. To Identify Key Predictors of College Dropout: To explore and identify a broad range of factors, including demographic, socioeconomic, academic, and financial variables, that significantly influence the likelihood of college students dropping out.
2. To Develop and Compare Predictive Models: To develop, train, and test various predictive models, including Logistic Regression, Random Forest, Decision Tree Classifier, Support Vector Machine (SVM), and Gradient Boosting, for their ability to predict college student dropout. This objective includes a comprehensive comparison of these models based on precision, recall, and F1 scores to determine the most effective model.
3. To Enhance Intervention Strategies: To leverage the insights gained from the predictive models to propose targeted intervention strategies for at-risk students, aiming to improve retention rates. This involves understanding the specific needs and challenges faced by different groups of students and designing interventions that address these issues effectively.
4. To Contribute to the Academic Literature: To provide a significant contribution to the academic literature on student retention by offering a detailed analysis of the predictive power of various models and the interplay of different factors affecting student dropout. This research aims to broaden the understanding of student retention challenges and the effectiveness of predictive modeling in addressing these issues.

Through these objectives, the research seeks to address the critical issue of college student dropout, offering practical tools and insights for educational institutions to improve student retention and success.

2.0 LITERATURE REVIEW

Student success in higher education is one that is of crucial importance to students, educators, and the general society as their success or failure have a ripple effect on everyone. As such several studies in the past have made significant strides in employing machine learning models and algorithms to predict student success and identify students at risk of dropping out. However, there is the need to identify robust machine learning and predictive models and factors that contribute to students dropping out.

A study by [3] contributed to reducing academic failure at higher education by using machine learning algorithms to identify students at risk. The study formulated a three-category classification task where there was a strong imbalance of the classes – success, relative success, and failure. It was found that boosting algorithms perform better than other models. Our study is related to that of Martins et al, but the current study focuses on a binary classification where we sampled the data to only students that have completed the credentials or dropped out excluding currently enrolled students. Also, the current study goes further to identify key factors and predictors that can potentially influence the likelihood of a college student drop out.

In a similar fashion, [2] meticulously utilized logistic regressions and decision trees to sift through examination data, aiming to discern patterns indicative of potential student dropouts. Their meticulous approach led to a notable achievement, with prediction accuracies soaring to an impressive 95% after just three semesters. Despite this commendable accuracy, the study's robustness came under scrutiny, as noted by [6] on comparison of just two predictive models. Concerns regarding the study's robustness prompt further investigation into the reliability and generalizability of the predictive models employed, emphasizing the need for comprehensive validation and sensitivity analyses to bolster the credibility of the findings.

Furthermore, [4] conducted a comparative analysis of various predictive models utilizing a dataset comprising 15,825 undergraduate students enrolled at the Budapest University of Technology and Economics between 2010 and 2017. Employing a robust methodological approach, they implemented 10-fold cross-validation to assess model performance. Their findings revealed that the area under the curve (AUC) for the best-performing models, gradient boosted trees, and deep learning, stood at 0.808 and 0.811, respectively. This compelling evidence also contributes to literature on the efficacy of gradient boosting and other boosting algorithms as superior predictive modeling approaches for anticipating student dropout rates. Such results underscore the importance of leveraging advanced machine learning techniques to enhance predictive accuracy and inform proactive intervention strategies in educational institutions.

3.0 METHODOLOGY

3.1 Data

The data employed in the research is from Polytechnic Institute of Portalegre. It has records of enrolled students between 2008/2009 – 2018/2019 academic year. The data was first used in a study of student success prediction by [3]. The data for the current study was sampled to focus on students that graduate and drop out. The variables include demographic information (age, gender, marital status, nationality), socioeconomic status (family income, parental education level), academic performance (grades, courses, enrolled credits), and financial factors (tuition payment, scholarship holder).

3.2 Exploratory Data Analysis

Before proceeding with predictive modelling, an exploratory data analysis (EDA) was conducted to assess the quality and structure of the collected data. This stage was crucial for identifying any inconsistencies, missing values, or outliers that could impact the accuracy of our models. The EDA process involved the following key steps:

1. Data Cleaning: The dataset was cleaned to address missing values, either by imputation or removal, depending on the extent and nature of the missing data. Inconsistent or duplicate records were corrected or eliminated to ensure the integrity of the analysis.
2. Descriptive Statistics: We computed descriptive statistics for all variables, including means, medians, standard deviations, and ranges. This provided an initial understanding of the data distribution and highlighted any potential anomalies or extreme values.

3. Correlation Analysis: A correlation analysis was conducted to examine the relationships between variables. This helped in identifying potential predictors for dropout and retention and in understanding the multicollinearity among the independent variables.

4. Visualization: Various visualization tools were employed, such as histograms, box plots, and scatter plots, to graphically represent the data distribution and relationships between variables. This facilitated the identification of patterns and trends that could inform our modeling approach.

The exploratory data analysis provided valuable insights into the dataset's characteristics and informed the subsequent phase of predictive modeling. By thoroughly understanding the data at hand, we were better equipped to select appropriate modeling techniques and to interpret the results of our analysis effectively. This foundational step was essential for ensuring the reliability and validity of our research findings.

3.3 Feature Engineering and Feature Selection

The data exploration helped identified features that are likely to have higher predictive power and features that may require some engineering to increase their predictive power. Target variable originally included dropout, graduate, and enrolled. The enrolled was dropped with focus on students that have graduated or dropped out to help create a binary prediction model. Moreover, the decision of enrolled students to graduate or drop out later in the future is unknown.

It was observed that age was not normally distributed, and this can affect the robustness of the model. Extreme or outliers can have more significant impact on the model estimates when the data is not normally distributed, potentially making predictions less reliable. The log of the age feature was taken to normalize the age feature. Logarithms reduces skewness and stabilize variance.

From the academic variables correlation analysis, it was noticed that first semester circular units approved, first semester circular units' grades, second semester circular units approved, and second semester circular units' grades are highly correlated with the target variable (graduate or dropout). Moreover, first semester circular units approved, and second semester circular units approved as well as first semester circular units' grade and second semester circular units graded are highly correlated. Therefore, the first semester circular units approved, and the first semester circular units' grades were selected as part of our predictive modelling. It must be noted that these variables were subjected to further evaluations before finally approve as part of the model.

Regarding correlation analysis for financial variables and target variable, it was noticed that tuition fees up-to-date, scholarship holder, and debtor variables are highly correlated with the target variable and less correlated among each other.

An initial logistic regression was from the selected variables in addition to the other non-dropped variables was run to identify the significance of these variables. The logic was to use p-values and R-square to remove variables that may potentially have less predictive power such that the model is simplified without significantly reducing its predictive performance.

3.4 Modelling

In building a predictive model to accurately predict students that are at risk of dropping out of college, the research built five predictive models – logistic regression, random forest, decision tree classifier, gradient boosting, and support vector machine. Several predictive models were employed for their benefits with small datasets as well as the imbalanced nature of the dataset. [5] stated that the reasoning process of decision tree models makes it easy to use and can be directly converted into set of IFTHEN rules. [1] employed Support Vector Machine for prediction as it works well with small datasets and has a good generalization ability. The model with the best precision, recall, and f1 score is selected. It must be noted that the target variable is unbalanced, and the use of accuracy can be misleading.

The data was split in 80 percent training and 20 percent testing. After which the models were trained on the training data and tested with the test data. The evaluation metrics were employed to see the performance of the predictive models. Furthermore, some hyper parameter tuning was done to improve model performance. Lastly, cross validation and evaluation were done to ascertain the generalized predictive power of the model.

4.0 RESULTS

4.1 Distribution of Variables

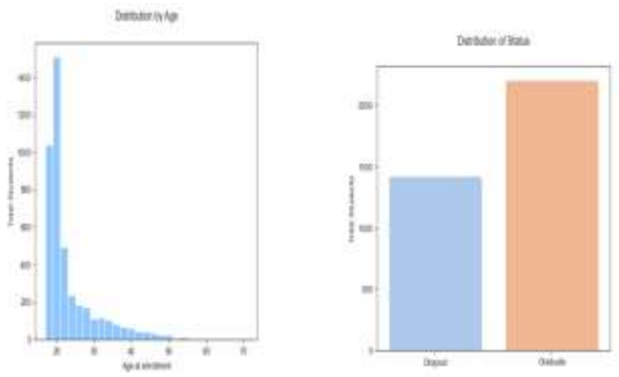

Figure 1: Distribution of age and target variable (status)

The age distribution shows that most of the students are between the ages of 18-21 and this is consistent with undergraduate enrolment. To use age in the predictive model, it is imperative to take the logarithm to make it normal. Moreover, the distribution of status or the target variable indicates that a lot more students graduate than drop out. However, it can be observed that the percentage of students that drop out is still high. Due to the imbalance nature of the target variable, accuracy alone may be misleading as an evaluation metric.

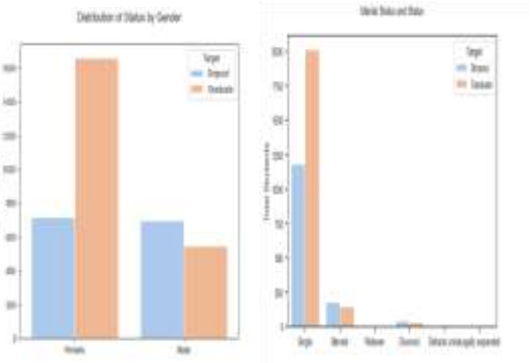

Figure 2: Distribution of Status by Gender and Marital Status

In the distribution of gender for the target variable, it can be observed aside from there being more female than male student, the drop out among male students is higher relative to female students. However, it is fairly the same among genders making gender to likely have a higher predictive power. It can be observed that a lot more students are single, and this is expected. However, due to a lot more students being single, marital status may not have a higher predictive power.

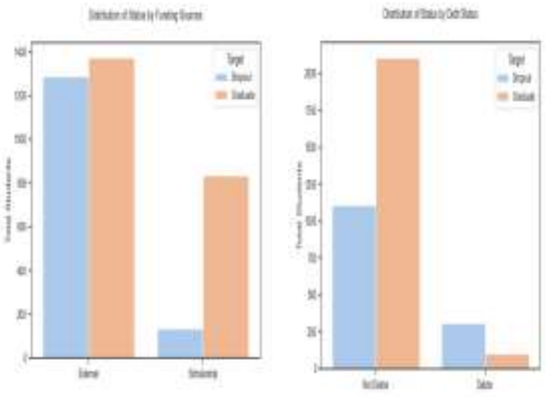

Figure 3: Distribution of Status by Funding Sources and Debt Status

It can be observed that students with scholarships have lower dropout rate relative to students that do not have scholarships. Also, dropout rate is higher among students categorized as debtors.

Table 1: Evaluation metrics results before cross validation and hyper parameter tuning

Evaluation Metrics	**Logistics**	**Decision Tree**	**Random Forest**	**Gradient Boosting**	**SVM**
Accuracy	0.718	0.861	0.875	0.877	0.873
Recall	0.923	0.938	0.944	0.942	0.858
Precision	0.705	0.852	0.865	0.870	0.953

F1 Score	0.799	0.893	0.903	0.905	0.903

Based on the f1 score, precision, and recall, gradient boosting stands as the best model to predict whether a student will drop out or graduate. Logistic regression model relatively had the lowest score among all the models, and this may be due to the imbalanced nature of the dataset.

Table 2: Evaluation metrics results after cross validation and hyper parameter tuning

Evaluation Metrics	**Logistics**	**Decision Tree**	**Random Forest**	**Gradient Boosting**	**SVM**
Average Accuracy	0.858	0.824	0.880	0.886	0.885
Average Recall	0.935	0.849	0.937	0.944	0.952
Average Precision	0.848	0.860	0.875	0.878	0.871
Average F1 Score	0.889	0.855	0.905	0.910	0.909

It can be observed from the results above that gradient boosting performs better than the other models when generally. An average recall of 0.944 means that on 94.4% of the students who dropped out were correctly identified by the model as dropouts. An average precision of 0.878 indicates that 87.8% of the students predicted to drop out by the model indeed dropped out. An average F1-score of 0.910 suggests that the model achieves a high balance between correctly identifying dropouts and avoiding false positives, demonstrating robustness in predicting both dropout and graduation outcomes. The results corroborate the finding of [3] that boosting methods make relatively better predictions for imbalanced data classifications. **4.2 Key Predictors in Gradient Boosting Model**

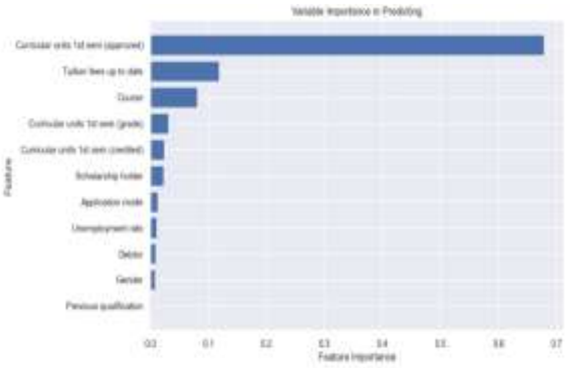

Figure 4: Strength of predictors used in gradient boosting model.

The above shows the order of importance of the variables employed in the predictive model to predict whether a student is likely to drop out or graduate. It can be observed that circular units approved in the first semester is the most important factor in determining student success in terms of graduating or dropping out. This is then followed by tuition fees up to date. The previous qualification of students has the lowest predictive power indicating that current circumstances of students – academic courses, and finances contributes top students' success more.

Future Work and Recommendations

Future work in the role of advanced data analytics in higher education should consider expanding the range of data sources. This includes incorporating behavioral data from Learning Management Systems (LMS) on student interactions and behaviors, socio-economic data such as financial aid and employment information, and psychological and social media data[7] on students' mental health, social interactions, and extracurricular activities.

Improved feature engineering is essential, focusing on developing domain-specific features that capture knowledge about subject difficulty, instructional quality, and pedagogical methods. Utilizing time-series analysis can help track and analyze changes in student performance and engagement over time. Applying Natural Language Processing (NLP) techniques to analyze textual data from student essays, discussion posts, and feedback can also provide valuable insights.

Deep learning models, including Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), should be considered for identifying complex data patterns. Implementing Explainable AI (XAI) techniques[8] can make model predictions more transparent and interpretable for educators and administrators.

Ethical considerations and bias mitigation are paramount. Regularly auditing models for biases related to race, gender, socio-economic status, and other protected characteristics is necessary. Ensuring robust data privacy and security measures to protect student information is essential. Adhering to ethical guidelines for the use of predictive analytics in education is crucial for maintaining trust and integrity[9].

Collaboration and stakeholder engagement are vital for the successful implementation of advanced data analytics in education. Encouraging interdisciplinary collaboration between data scientists, educators, psychologists, and sociologists can lead to more holistic solutions. Engaging students, faculty, and administrators in the design and implementation of predictive analytics solutions ensures that the systems are user-centered and effective. Working with policymakers to develop guidelines and policies that support the ethical and effective use of advanced data analytics in education is also recommended.

5.0 CONCLUSION

In conclusion, this research delves into the critical issue of college student dropout, employing advanced predictive modeling techniques to forecast students' likelihood of discontinuing their studies. By leveraging a comprehensive dataset encompassing demographic, socioeconomic, academic, and financial factors, our analysis underscores the complexity of student retention and the necessity for nuanced approaches to address it.

Our findings demonstrate that the gradient boosting model outperforms other models in predicting student dropout, achieving high precision, recall, and F1 scores. This underscores the model's robustness in handling the multifaceted nature of factors influencing college retention. Through targeted interventions informed by predictive modeling, educational institutions can enhance retention rates and contribute to broader objectives of educational equity and success.

In an era marked by rapid changes in education and workforce dynamics, our research holds significant implications. By identifying at risk students, institutions can tailor interventions to address specific challenges, fostering a more supportive and inclusive educational environment. Ultimately, this work contributes to both practical strategies for improving retention and the theoretical understanding of student persistence in higher education.

REFERENCES

[1] Hämäläinen, W., & Vinni, M. (2006, June). Comparison of machine learning methods for intelligent tutoring systems. In International conference on intelligent tutoring systems (pp. 525-534). Berlin, Heidelberg: Springer Berlin Heidelberg.

[2] Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. European Journal of Higher Education, 10(1), 28-47.

[3] Martins, M. V., Tolledo, D., Machado, J., Baptista, L. M., & Realinho, V. (2021). Early prediction of student's performance in higher education: A case study. In Trends and Applications in Information Systems and Technologies: Volume 1 9 (pp. 166175). Springer International Publishing.

[4] Nagy, M., & Molontay, R. (2018). Predicting dropout in higher education based on secondary school performance. In 2018 IEEE 22nd international conference on intelligent engineering systems (INES) (pp. 000389-000394). IEEE.

[5] Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008). Data mining algorithms to classify students. In Educational data mining 2008

[6] Singh, A., Saraswat, S., & Faujdar, N. (2017, May). Analyzing Titanic disaster using machine learning algorithms. In 2017 International Conference on

Computing, Communication and Automation (ICCCA) (pp. 406-411). IEEE.

[7] Olola, Toyosi & Akpan, Ubong-Abasi Beatitudes & Odufuwa, Funmilayo. (2022). Investigation of the Psychological Effects of Social Media Use Among Students in Minnesota, United State America. International Journal of International Relations, Media and Mass Communication Studies. 8. 37-47. 10.37745/ijirmmcs.15/vol8n33747.

[8] Chiamaka Daniella Okenwa & Omoyin Damilola. David & Adeyinka Orelaja. & Oladayo Tosin Akinwande, 2024. "Exploring the Role of Explainable AI in Compliance Models for Fraud Prevention," International Journal of Latest Technology in Engineering, Management & Applied Science, International Journal of Latest Technology in Engineering, Management & Applied Science (IJLTEMAS), vol. 13(5), pages 232-239, May.

[9] Ifeoluwa Oladele, Adeyinka Orelaja and Oladayo Tosin Akinwande, 2004. "Ethical Implications and Governance of Artificial Intelligence in Business Decisions: A Deep Dive into the Ethical Challenges and Governance Issues Surrounding the Use of Artificial Intelligence in Making Critical Business Decisions" International Journal of Latest Technology in Engineering, Management & Applied Science-IJLTEMAS vol.13 issue 2, February 2024, pp.48-56 URL: https://doi.org/10.51583/IJLTEMAS.2024.130207