IS 804
Project Proposal
Kelvin Echenim
FY95131
March 27, 2023

Predicting Academic Success and Retention Rates of Higher Education Students

ABSTRACT

This research study aims to predict academic success and dropout rates among college students. The study aims to examine the predictors made available on a publicly accessible Kaggle dataset, which offers detailed data on students enrolled in different undergraduate degrees. The dataset contains information relating to academic data, macroeconomic factors, social-economic factors, and demographics that are used in this study to investigate potential predictors of student dropout and academic success. Additionally, the data provide illuminating details about the factors that affect student success, which higher education institutions may use to guide interventions and rules on student retention. The resultant information could help higher education institutions objectively understand how their students progress academically and identify areas requiring personal and institutional improvement. When used, the predictive models created in this project will consider the student dropout risk factors discovered from the dataset, encouraging early intervention to boost retention rates. As a result, they can develop academic initiatives, programs, or activities specifically targeted at enhancing academic achievement.

INTRODUCTION

To ensure the success of undergraduates in school and life after graduation, higher education institutions must be responsible for supporting their academic and social development. However, some students still need help to finish their degree programs, despite the concerted efforts of higher education institutions to provide the essential tools for student success. Recent studies show that despite efforts to address the issue, student dropout rates in higher education have stayed relatively constant over the past ten years. Dropout rates have a wide range of negative consequences, including financial loss for institutions and students and a reduction in the human capital of the workforce.

Therefore, it is essential in higher education to forecast academic success and dropout rates. This study examines the factors influencing college students' dropout rates and academic success. The research will use a publicly accessible Kaggle dataset that offers comprehensive information on students enrolled in various undergraduate degrees. Investigating potential predictors of student dropout and academic success can be done using the dataset's information on academics, macroeconomics and socioeconomic factors, and demographics. Each record corresponds to a unique student, with 35 features and 4424 data points.

The development of predictive models that can recognize student dropout risk factors is the main objective of this study, as doing so will allow for early interventions that will increase retention rates. Additionally, the study will offer insightful information on the variables that affect student success, information that higher education institutions may use to direct interventions, and policies on student retention. The institutions will be able to more objectively understand how their students progress academically and identify any areas that need individual and institutional improvement with the help of the information gathered from this study.

The contribution of this work is in its potential to support higher education institutions in creating academic initiatives, programs, or activities explicitly aimed at improving academic performance. Higher education institutions can improve student retention rates and overall academic success by taking proactive measures to identify the factors that predict academic success and dropout rates.

LITERATURE REVIEW

Predicting students' academic success and dropout rates is an antiquated concept and an ongoing concern in higher education. In order to increase retention rates and student success, researchers have been looking into various predictors of academic success and dropout rates. Numerous studies have identified metrics and developed methodologies to objectively forecast how well a student will perform in a structured educational setting. According to Pea-Ayala Alejandro (2014), scholars and academic administrators have been keen on predicting student success in academic institutions for the last two decades. This multidisciplinary research will continue as predictive models, and data mining evolve in their algorithms.

This study is critical because it elaborates on the advancements made in ensuring student retention and success are essential to the mission of higher education institutions. According to Tight et al. (2020), academic institutions worldwide face challenges with student retention due to a lack of funding and resources, with an average dropout rate of about 45% in the countries that make up the Organization for Economic Co-operation and Development (OECD). In order to address this problem, higher education institutions are continually developing and implementing intervention strategies. However, professionals and academics agree that these techniques peak during the first year of study. As a result, it has received much attention to identifying vulnerable students likely to drop out of their courses as soon as possible (Zeineddine et al. 2021).

Several studies have shown that academic performance, socioeconomic factors, and demographic factors are some of the most important predictors of college student's academic success and dropout rates. For instance, Carlos Felipe Rodrguez-Hernández's (2020) systematic findings revealed that socioeconomic factors and academic performance significantly predict student dropout rates. Similarly, Salvatore A. Barbera et al. (2020) found that demographic factors like race/ethnicity and gender also impact student retention rates after reviewing undergraduate student retention and graduation since 2010.

Despite the findings of these studies, further research is still needed to identify additional predictors of academic success and dropout rates due to the far-reaching effects of student dropouts vis-a-vis the changing times. One way to address this gap is by using data mining techniques to explore large, recent datasets and identify potential predictors researchers in previous studies may have overlooked. This study will use predictive modeling and data mining techniques to explore a publicly accessible Kaggle dataset to identify potential predictors of academic success and college dropout rates. It is a massive, heterogeneous dataset with characteristics that could yield novel insights in this area of research. We hypothesize that student demographics, macroeconomic factors, socioeconomic status, and academic performance will significantly predict college students' academic success and dropout rates

AIM AND OBJECTIVES

This study aims to predict academic success and dropout rates among college students using a Kaggle dataset that is accessible to the public. The study will identify the key predictors of academic success and student dropout rates, enabling higher education institutions to take proactive measures to improve student retention rates.

We will pursue the following objectives (using machine learning techniques) to achieve the aim of the study: First, estimating the most significant antecedents of academic success and dropout rates among college students (based on the available dataset) and second, developing predictive models that can identify risk factors for student dropout, allowing for early intervention to increase retention rates.

Subsequently, this study provides illuminating details about the variables that considerably influence student success, which higher education institutions can use as guidelines for interventions and student retention. By achieving these objectives, the study will contribute to the existing body of knowledge by identifying additional predictors of academic success and dropout rates and developing predictive models to identify student dropout risk factors, allowing for early intervention to increase retention rates, which corroborates their business case.

RESEARCH QUESTIONS/HYPOTHESES

The following research questions arise:
What are the key predictors of academic success among undergraduate students?
What are the key predictors of student dropout rates among undergraduate students?
How can predictive models be developed to identify student dropout risk factors and encourage early intervention to boost retention rates?

The associated hypotheses are as follows:
Hypothesis 1: Academic grades/performance will significantly predict academic success among undergraduate students.
Hypothesis 2: Socio-economic factors such as Scholarship holder, Parental qualification, Parental occupation, Debtor, Tuition fees up-to-date, and Educational special needs will significantly predict academic success among undergraduate students.
Hypothesis 3: Demographic factors such as Displaced, Gender, Nationality, Marital status, Age at enrollment, and International will be significant predictors of academic success among undergraduate students, opening doors for targeted assistance and closing accessibility gaps.
Hypothesis 4: Macroeconomic factors such as the Unemployment rate, Inflation rate, and Gross Domestic Product (GDP) will significantly predict academic success among undergraduate students.
Hypothesis 5: Predictive models developed using data mining techniques will be able to identify student dropout risk factors accurately.

METHODS

Theoretical Basis:
The methods used in this study are based on the theoretical framework of predictive analytics and data mining techniques. Predictive analytics uses data, statistical algorithms, and machine learning techniques to determine the likelihood of future outcomes based on historical data. Data mining is analyzing large datasets to discover patterns, trends, and relationships that can be used to make informed decisions. The R programming language will be employed in the ensuing analysis.

Population and Sample:
The population for this study is college students enrolled in different undergraduate degrees in Europe. The Sample will be drawn from the publicly accessible Kaggle dataset, which contains detailed information on students' academic data, macroeconomic and socio-economic factors, and demographics. The dataset includes approximately 4424 students from 17 colleges between 2008/2009 and 2018/2019.

Justification for Sample size:
The sample size for this study was determined based on the size of the Kaggle dataset and the availability of relevant data. Therefore, a sample size of approximately 4424 students, with 35 features each, was deemed appropriate for the analysis.

Procedures for Data Collection:
The data for this study will be collected from the publicly accessible Kaggle dataset. The dataset contains students' academic data, macroeconomic factors, socio-economic factors, and demographics. We will pre-process the data to ensure that the data is accurate and complete for subsequent analysis.

Procedures to Address Validity and Reliability:
The following procedures will be implemented in order to address the validity and reliability

of the data:

Data Pre-processing:
The data will be pre-processed to ensure that the data is accurate and complete. Any missing data will be imputed using appropriate imputation techniques.

Sampling Techniques:
Since the dataset contained unbalanced classes, methods such as Synthetic Minority Oversampling Technique (SMOTE) will be utilized to ensure the classification algorithm does not unduly favor the majority class.

Feature Extraction:
Feature extraction methods will be employed. The collinearity of the features will also be used to select relevant attributes.

Statistical Analysis:
Appropriate statistical techniques, including regression and machine learning techniques, will be used to analyze the data.

Cross-Validation:
We will use cross-validation techniques to ensure the models are accurate and reliable, serving to test the predictive models' validity.

By adhering to these procedures, the study will ensure the validity and reliability of the data and the results obtained from the analysis.

EXPECTED OUTCOMES

The study aims to accomplish the following goals:

Identify significant predictors of academic success and dropout rates among college students:
The study will examine the predictors available on the Kaggle dataset, including academic data, macroeconomic and socioeconomic factors, and demographics. The study will identify the most significant predictors of academic success and dropout rates among college students.

Develop predictive models for academic success and dropout rates using the identified predictors:
The models will employ machine learning algorithms (developed in the R programming language) to recognize patterns and trends we will use to predict student outcomes.

Provide higher education institutions with insights to guide interventions and rules on student retention:
The study's findings will provide higher education institutions with insights to guide interventions and retention policies. In addition, the findings could assist institutions in objectively

assessing their student's academic progress and identifying areas requiring personal and institutional improvement.

Encourage early intervention to boost retention rates:
The predictive models developed as part of this project will consider the student dropout risk factors identified from the dataset, encouraging early intervention to increase retention rates. Consequently, higher learning institutions can develop academic initiatives, programs, or activities to enhance academic achievement.

Overall, the anticipated outcomes of this study will contribute to a better understanding of the factors that influence student success and dropout rates in colleges and provide higher education institutions with insight into how to improve student retention rates.

PROJECT MANAGEMENT

Research Personnel:
The study is a solo, class research project which will foster the application of information science and research concepts learned in class. The research personnel, a doctoral student in the Information Systems department, under the supervision of the Professor in charge of the course, will conduct the data analysis, development of the predictive models, and interpretation of the results.

Timeline of Research:
Beginning on the date of project approval in February 2023, the Research will be conducted over three months. The timeline for the study will be as follows:
Weeks 1–4: Data collection and cleaning
Weeks 5–6: Project Proposal Writing
Weeks 7-8: Exploratory data analysis and feature engineering
Weeks 9–10: Predictive model development and validation
Weeks 11-12: Results interpretation and report writing
Week 12: Submission of findings and final report

Evaluation Plan:
The Research will be evaluated based on the following criteria:
Rigor and validity of the data analysis methods
The robustness, precision, recall, and accuracy of the developed predictive models
Clarity and relevance of the study findings
Impact of the study on higher education institutions' understanding of the factors affecting student success and dropout rates.
We will employ established data analysis techniques and machine learning algorithms to ensure the validity and dependability of the study's findings. Also, we will use an integrated development environment for the R programming language to run the codes.

RESEARCH JUSTIFICATION

The research study aims to predict college students' academic success and dropout rates by examining the available predictors on a publicly accessible Kaggle dataset. The study will use machine learning algorithms to develop predictive models to identify student dropout and academic success risk factors.

The intellectual merit of this study lies in its potential to advance knowledge in the field of education research, particularly in student retention and academic success. More so, the study will provide a comprehensive understanding of the factors that affect student success and dropout rates, allowing higher education institutions to develop targeted interventions and programs to enhance student success.

The broader impacts of this study include its potential to benefit society by improving the graduation rates of college students. This research will provide insights into how higher education institutions can better support their students, especially those at risk of dropping out. The study findings could assist colleges and universities in designing retention and academic success-promoting policies and programs.

The beneficiaries of this research include higher education institutions, policymakers, and college students. Institutions of higher education can utilize the findings of this study to develop policies and practices based on empirical evidence that promote student success. Policymakers can leverage the findings of this research to develop policies that promote student success and improve the overall quality of education. Ultimately, the increased support and resources that higher education institutions can provide due to this research will benefit college students.