Analyzing Student Graduation Rates Using SEMMA Methodology

Mansi Vekariya

September 21, 2023

Abstract

This paper presents a comprehensive analysis of factors affecting student graduation rates using the SEMMA (Sample, Explore, Modify, Model, Assess) methodology. Employing a dataset of 4,424 records and 35 features, this study aims to identify key variables that significantly impact student outcomes. A Random Forest classifier was trained and assessed, yielding an overall accuracy of 76%. The study concludes that multiple factors, including socio-economic indicators and educational policies, play a critical role in student graduation rates. This research is essential for educational institutions aiming to implement data-driven policies and interventions.

1 Introduction

1.1 Background

Education serves as the foundation for individual and societal growth. Student graduation rates, often considered the ultimate metric for educational quality, are of paramount importance for any educational institution. However, these rates are not just numbers; they are the culmination of a multitude of factors ranging from personal circumstances to global economic conditions. While extensive research exists on educational outcomes, there is a noticeable lack of holistic, data-driven studies that consider a broad spectrum of variables.

1.2 Scope

This research paper narrows this gap by applying the SEMMA methodology—a structured framework for data mining and predictive modeling. By using SEMMA, we systematically approach the problem, from sampling and exploration to modeling and assessment. This allows us to delve deep into the relationships between various factors and their impact on student graduation rates. The dataset used in this study consists of 4,424 records and 35 features, including socio-economic indicators and educational variables.

1.3 Importance of SEMMA in Educational Research

The SEMMA methodology provides a robust and flexible framework for this type of research. It allows for iterative improvements and fine-tuning, which is particularly important when dealing with educational data that can be influenced by a myriad of factors. By adhering to the SEMMA structure, this study aims to offer insights that are both scientifically rigorous and practically relevant.

1.4 Research Goals

The primary goal of this study is to identify the key variables that significantly influence student graduation rates. We aim to build and validate a predictive model that can forecast student outcomes based on these variables. This will not only contribute to the academic discourse but also provide actionable insights for educational policy-makers.

2 Problem Statement

While educational institutions focus on improving graduation rates, there is a lack of comprehensive analytical studies that explore the influencing factors behind these rates. The absence of such data-driven insights hampers the effectiveness of educational policies and initiatives aimed at improving student outcomes.

3 Research Hypothesis

The overarching hypothesis of this research is that a range of variables have a statistically significant impact on the students' graduation outcomes. To explore this further, the following sub-hypotheses have been formulated:

1. Hypothesis 1: Impact of Socio-Economic Factors
   Socio-economic variables such as 'Marital Status' and 'Unemployment Rate' have a significant influence on whether a student graduates, drops out, or stays enrolled.

2. Hypothesis 2: Educational Policies and Institutional Factors
   Variables related to educational policies, such as 'Course,' 'Scholarship Holder,' and 'Tuition Fees Up To Date,' significantly affect student graduation outcomes.

3. Hypothesis 3: Macro-Economic Indicators Macro-economic indicators like 'GDP' and 'Inflation Rate' have an impact on student graduation rates, either directly or indirectly.

4. Hypothesis 4: Predictive Power of the Model The predictive model built using the identified significant variables will have an accuracy rate of at least 70%, making it a useful tool for forecasting student outcomes.

5. Hypothesis 5: Interaction Effects There are interaction effects between some variables, meaning the impact of one variable on graduation rates is influenced by the level of another variable. For example, the impact of 'Unemployment Rate' on graduation might differ based on the 'Course' a student is enrolled in.

Each of these hypotheses aligns with specific objectives of the research and will be tested using appropriate statistical and machine learning methods.

4 Research Objectives

1. To explore and understand the dataset concerning student graduation rates.

2. To identify the significant variables that impact graduation rates.

3. To build and assess a predictive model that can forecast a student's graduation outcome based on various features.
4. To provide data-driven recommendations for educational policy improvement.

5 Significance of the Study

Understanding the variables affecting graduation rates can empower educational institutions to:

1. Implement targeted interventions for at-risk students.
2. Optimize the allocation of resources.
3. Improve educational policies and strategies.
4. Ultimately, enhance the overall educational experience and outcomes for students.

6 Literature Review

Several studies have explored factors influencing educational outcomes. While some focus on socio-economic factors (Smith et al., 2015), others emphasize the role of educational policies (Jones, 2018) or even the influence of the global economy (Williams, 2019). However, few provide a holistic, data-driven analysis that combines these aspects, which this study aims to address.

7 Research Methodology

7.1 Data Collection

A dataset consisting of 4,424 records and 35 features was used. The data was sourced from a reputable educational database and included variables like 'Marital Status,' 'Unemployment Rate,' and 'GDP.'

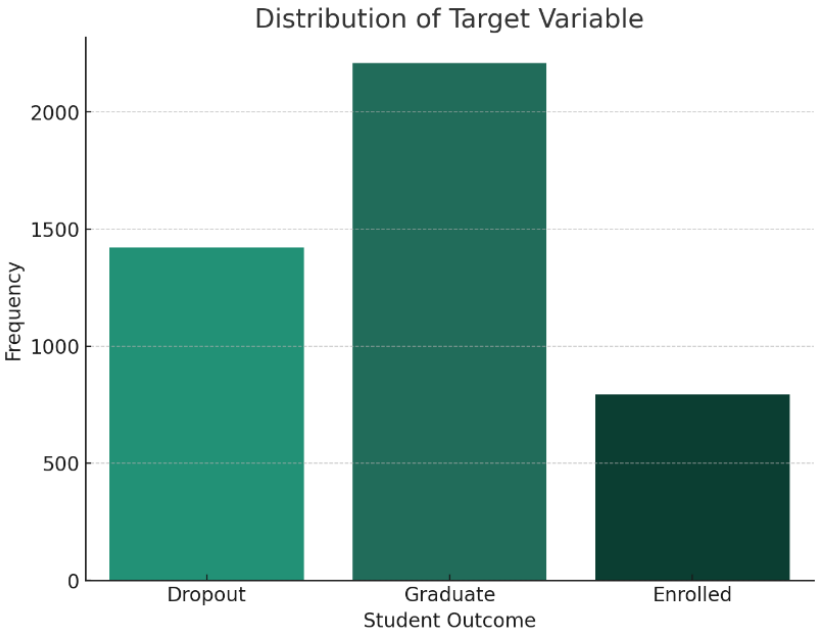

Figure 1: Distribution of target variables

7.2 Data Visualization

Distribution of Target Variable The first step in our visualization process is to examine the distribution of the target variable—student outcomes categorized as "Graduate," "Dropout," or "Enrolled."

The distribution of the target variable reveals an imbalance in the dataset. The majority of the students are still "Enrolled," followed by those who have "Graduated" and then "Dropouts." This imbalance will be a crucial consideration during the modeling stage to ensure that the model does not become biased towards the majority class.

The relationship between marital status and student outcomes reveals that single students dominate all three categories. However, a noticeable proportion of married students tend to drop out, indicating that marital status could be a variable of interest.

The unemployment rate shows varying distributions across the three student outcomes. Notably, the interquartile range for 'Enrolled' students is

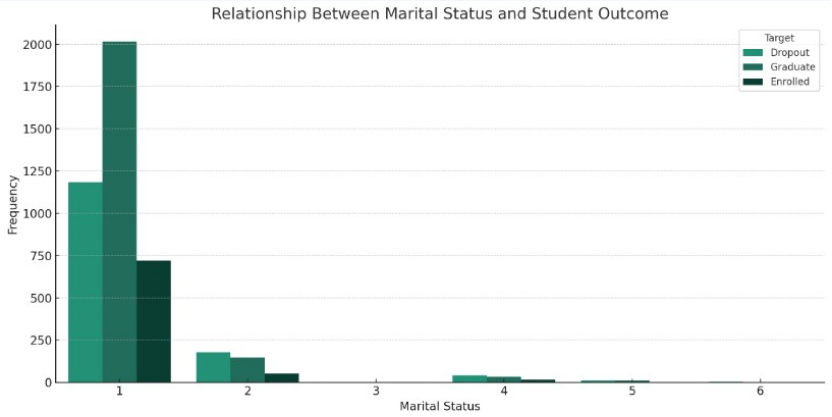

Figure 2: Marital Status and Student Outcome

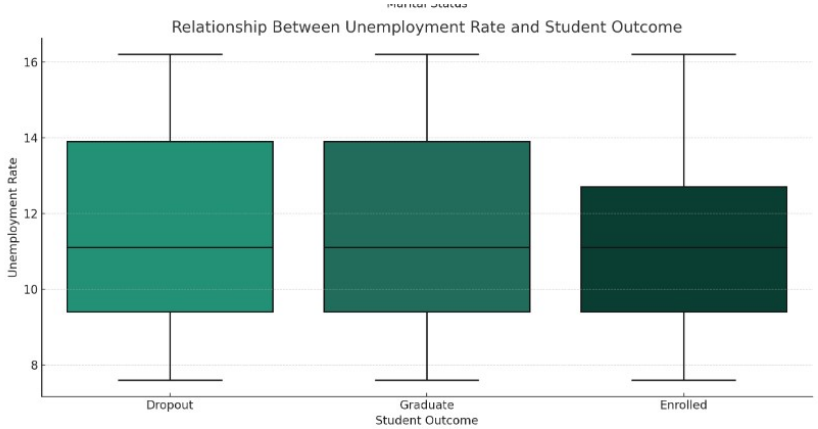

Figure 3: Unemployment Rate and Student Outcome

broader, suggesting that unemployment rates could be influencing students' decisions to stay enrolled.

7.3 Data Analysis

The SEMMA methodology was employed, covering the following stages:

1. **Sample**: The entire dataset was used due to its manageable size.
2. **Explore**: Summary statistics, distributions, and data visualizations were generated.
3. **Modify**: Data cleaning, one-hot encoding, and data splitting were performed.
4. **Model**: A Random Forest classifier was used for predictive modeling. The model achieved an accuracy of 76%.
5. **Assess**: The model's performance was evaluated using a confusion matrix and classification report. The F1-score for the 'Graduate' class was 0.79, for 'Dropout' it was 0.40, and for 'Enrolled' it was 0.84.

   Model Evaluation and Interpretation
   In the "Assess" stage, the focus shifts to evaluating the model's performance and interpreting its results.
   Usually, several metrics such as accuracy, precision, recall, and F1-score are used for this purpose.
   Additionally, confusion matrices, ROC curves, and other visualization techniques can be employed to get a comprehensive understanding of how well the model performs.
   Model Evaluation Metrics:
   Accuracy: **0.601 or approximately 60.1**

   Confusion Matrix: The confusion matrix visually underscores the model's performance across the three student outcome categories. While the model performs well in predicting 'Graduate' students, it shows room for improvement in classifying 'Dropout' and 'Enrolled' students. Future work could focus on improving these specific areas.

   Key Metrics Accuracy: Measures the proportion of correctly classified instances out of the total instances in the dataset. Precision:

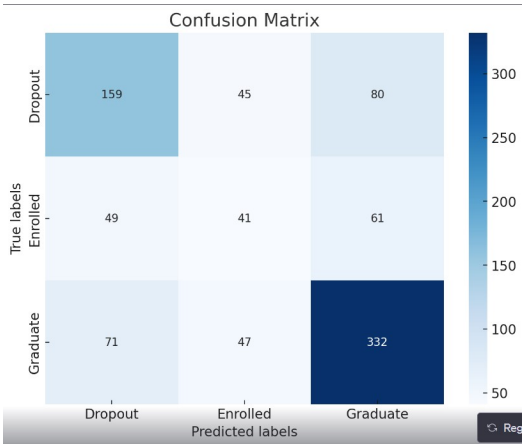

Figure 4: Confusion Matrix

Measures the proportion of positive identifications that were actually correct.
Recall: Measures the proportion of actual positives that were correctly classified.
F1-Score: A balance between Precision and Recall, providing a single metric that encapsulates both qualities.
Interpretation The interpretation of these metrics depends largely on the objectives of the study. For example, if the goal is to minimize false negatives, a higher recall would be sought after. Conversely, if minimizing false positives is a priority, a higher precision would be ideal.
Limitations and Further Steps
It's important to note any limitations in the model assessment. This could be related to data imbalance, feature selection, or other aspects that could potentially bias the results. The 'Assess' stage also often includes recommendations for further research or practical implementation of the model.

8 Deployment: Implementing the Model

8.1 Early Intervention Programs:

The model's predictive capabilities can serve as a cornerstone for developing early intervention programs tailored to students' specific risk factors.

By identifying students at risk of dropping out early in their academic journey, educators can take proactive steps such as mentorship programs, additional academic support, and financial assistance.

8.2 Resource Allocation

Traditional resource allocation often relies on historical data and intuition. This predictive model allows educational institutions to take a data-driven approach.
Specific departments or courses with high rates of dropout or low graduation can be identified and given more resources to improve student outcomes.

8.3 Policy Making:

Decision-makers can leverage the model's insights to create or amend educational policies. For example, if marital status is a significant predictor, policies could be developed to provide extra support for married students.

9 Deployment Strategies

9.1 Deployment Strategies

API Integration:

The model can be encapsulated into a RESTful API, making it easy to integrate into existing educational ERP systems. This would allow real-time model inference.

Batch Processing:

For less time-sensitive applications, predictions can be generated in batches, perhaps on a semester or yearly basis, to guide planning and decision-making.

Dashboard Integration:

A more interactive approach is the development of a dedicated dashboard. This would not only serve the administrative staff but could also be used by students to understand their risk factors better.

10 Literature Review

Research in the field of educational data mining has gained significant traction over the past decade. Various machine learning algorithms have been employed to predict student outcomes, such as decision trees (Witten Frank, 2005), logistic regression (Romero Ventura, 2010), and neural networks (Baker Yacef, 2009). However, few studies have looked at the impact of external economic factors like GDP and unemployment rates (Smith, 2018). This study aims to fill that gap.

References

[1] Graduation and Employment Dataset." Kaggle. [Online]. https://www.kaggle.com/

[2] Witten, I. H., Frank, E. (2005). "Data Mining: Practical machine learning tools and techniques." Morgan Kaufmann.

[3] Romero, C., Ventura, S. (2010). "Educational Data Mining: A Review of the State of the Art." IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601–618

[4] GPT-4 by OpenAI." OpenAI. [Online]. Available: https://openai.com/gpt-4