**Student Academic Dropout and Success: Investigating the impact of Demographic, Social and Economic Factors**

Team: Adrian Alexander, Shaun McKellar, Victoria Zangen

IST 652 Spring 2024 Final Project

Source:

<https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data>

About the dataset:

The original dataset provides a snapshot of students enrolled in a foreign university’s undergraduate degree program in the form of a csv file containing 35 columns and 4,424 rows. Each row represents an in-depth overview of one student in their program of choice, various demographic and socioeconomic factors, and their educational outcome.

By including variables such as unemployment rate, inflation rate, and GDP from their respective regions; the university can understand what critical economic data or external factors may also influence the student’s decision-making process.

Analyzing the data can assist in predicting a student performance level by assessing the units credited/enrolled/evaluated/approved, their grade each term, and the success or failure rate.

Understanding the data:

The raw data we call student\_data is comprised of 35 attributes:

* Marital Status: Marital status of the student
* Application Mode: Method of application used
* Course: The course taken by the student
* Daytime/Evening Attendance: Student attends Day or Evening class
* Previous Qualification: Qualification before enrolling in higher ed
* Nationality: Nationality of the student
* Mother/Father Qualification: Qualification of the mother/father
* Mother/Father Occupation: Occupation of mother/father
* Displaced: Whether the student is a displaced person
* Educational Special Needs: Student have any special educational needs
* Debtor: Is the student a debtor
* Tuition Fees Up to Date: Payment status
* Gender: Gender of student
* Scholarship Holder: Is the student a scholarship holder
* International: Is the student an international student
* Application Order: The order in which the student applied
* Age at Enrollment: The age of the student at the time of enrollment
* Curricular units 1st/2nd Sem (credited): Number of credited units
* Curricular units 1st/2nd Sem (enrolled): Number of enrolled units
* Curricular units 1st/2nd Sem (evaluations): Number of evaluated units
* Curricular units 1st/2nd Sem (approved): Number of approved units
* Curricular units 1st/2nd Sem (grade): Grade Units
* Curricular units 1st/2nd Sem (Without Evaluations): Number of units not evaluated
* Unemployment Rate
* Inflation Rate
* GDP
* Target

Cleaning the Data:

We saved the raw data student\_data to a new dataframe and called it originalstudent\_df in order to explore and understand how the data needed to be cleaned. Originalstudent\_df did not have any missing data, and the data types aligned with each attribute. We did note the ‘Target’ attribute was defined as an object using ‘Dropout’, ‘Enrolled’ and ‘Graduate’ in its values. Using scikit-learn, we recoded this attribute to use 0,1,2 categorical values instead.

Example:

Originalstudent\_df cleanstudent\_df

A screenshot of a graph

Description automatically generatedA table with numbers and letters

Description automatically generatedA screenshot of a white table

Description automatically generatedA screenshot of a graph

Description automatically generated

Methods of Analysis:

* Descriptive statistics by attribute
  + Min/Max, Mean, Median, Mode
    - Analysis Questions: What does descriptive statistics tell us about the data?
* Inferential Statistics
  + Linear Regression
  + Correlation
    - Analysis Questions: Are there relationships between the target academic success or dropout with demographic, social-economics, academic performance or economic factors?
* Predictive Analysis
  + Supervised Machine Learning
    - Analysis Questions: Can student dropout or academic success be predicted?
* Prescriptive Analysis
  + - Analysis Questions: How can this data help educational institutions in the future?

We intend our project to be larger in scope than the homework by considering predictive factors that may influence or contribute to a student’s retention or dropout rate. Throughout our analysis, it will also be important to understand any biases or risks present in the data.

Descriptive Statistics:

A group of colorful text

Description automatically generated with medium confidenceIn order to understand the make up of our data, we used descriptive statistics using .describe(), bar graphs to show the attribute value frequencies, and box plots. Tables describing the attribute value definitions can be found at the end of the report.

A screenshot of a graph

Description automatically generated

A chart of a diagram

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generated

**A pie chart with numbers and a blue circle

Description automatically generatedWhat does descriptive statistics tell us about the data?**

It is important to understand the target distribution of the dataset Dropout (0), Enrolled (1), Graduate (2). The initial finding for this semester’s data notes almost 50% of students graduate, 33% drop out and 18% are enrolled. This information is the starting point to understanding what contributing factors may play a role in a student decision to drop out, or graduate.

Based on the breakdown of information collected through descriptive statistics, several variables may influence the student graduation rate at the University. In order to understand which key attributes, play a significant role, we broke the variables down into five categories.

A close-up of a computer screen

Description automatically generated

The last step in cleaning the data before any further analysis was to subset cleanstudent\_df into cleanstudent\_df (students who graduate and drop out) and enrolledstudent\_df (students who are enrolled). The purpose is to analyze the data collected for students who graduate or drop out, and use this information in a supervised machine learning model to predict the students success rate in enrolledstudent\_df.

Number of students:

cleanstudent\_df: 3630 enrolledstudent\_df: 794

Running the descriptive statistics again on our cleanstudent\_df we find:

A blue and orange pie chart

Description automatically generated

Inferential Statistics:

A diagram of a graph

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generatedCorrelations help to determine what attributes play a significant role in a students decision to dropout or continue their studies.

Looking at the correlation values in relation to the Target attribute will help determine the student’s academic outcome. By focusing on the most correlated attributes less than -.2 and greater than .2 this will remove attributes that are not meaningful or significant in our analysis. We called this new dataframe cleanstudent\_subset.

The top 8 correlated attributes by category are:

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

As a group we utilized multiple linear regression models to determine the relationships between the demographic socioeconomic, and academic performance variables against the target/student enrollment status. We found interesting results through our models. Our model on socioeconomic variables against target resulted in a R squared value of 0.25. This means that 25% of the change in student dropout status is explained by these variables. Our model on Academic Performance variables against target resulted in a R squared value of 0.45. This means that 45% of the change in student dropout status is explained by these variables. Our model on Demographic variables against target resulted in a R squared value of 0.11. This means that 11% of the change in student dropout status is explained by these variables.

**Are there relationships between the target academic success or dropout with demographic, social-economics, academic performance or economic factors?**

* It seems that attributes that are categorized under Academic Performance are the most influential in a student’s dropout or graduation rate.
* Intuitively a student’s grade may place them on probation or at higher risk of dropping out if they do not perform well.
* Socio-Economics play a minor role in a student’s success or dropout rate.
* Students who are debt free and up to date on their tuition payments, will more than likely graduate. While students who have financial troubles may drop out.
* Age at enrollment and gender have little significance towards a student’s dropout or success rate.
* Academic enrollment and economic factors do not contribute towards student’s dropout or graduation rate.

Predictive Analysis:

To predict student outcomes more accurately, we employed supervised machine learning, specifically a Random Forest algorithm. Random Forest is ideal for our dataset due to its capability to handle numerous features and its robustness against overfitting.

Running the Random Forest model on the originalstudent\_df achieved an accuracy of 77%.

The confusion matrix shows:

* True Positives: 594 Correct predictions of students who graduated.
* True Negatives: 352 Correct predictions of students who dropped out.
* False Positives: 63 Students incorrectly predicted to graduate.
* False Negatives: 19 Students incorrectly predicted to dop out.

This matrix not only reflects the model’s accuracy but also its reliability in predicting different student outcomes. We ran the same model training on cleanstudent\_subset and achieved an accuracy of 88%.

The confusion matrix shows:

* True Positives: 631 Correct predictions of students who graduated.
* True Negatives: 332 Correct predictions of students who dropped out.
* False Positives: 82 Students incorrectly predicted to graduate.
* False Negatives: 44 Students incorrectly predicted to dop out.

By focusing on the top 8 attributes for incoming cohorts, the university ca help boost their graduation rates in the future.

Using the training model from cleanstudent\_subset, we can apply these learning to test the number of students who are currently enrolled in the university to understand the chance of a student dropping out or graduating in the future. Running this model on enrolledstudent\_subset we find:

* True Positives: 432 Correct predictions of students who graduated.
* True Negatives: 362 Correct predictions of students who dropped out.

**Can student dropout or academic success be predicted?**

Identifying the most relevant variables in the dataset, and collecting information regularly will improve the training model creating a more accurate model in predicting student drop out or graduation rates.

Prescriptive Analysis:

**How can this data help educational institutions in the future?**

* High drop out rates may suggest potential issues in the program
* Early intervention for students who are at a higher risk of dropping out can be supported with services like success advisors, peer support groups, or financial aid advisors to boost program reputations and rankings
* Graduation rates influence prospective students and foster a positive reputation for the institution
* Student dropout rates may target specific demographic, socio-economic factors, or ethnic backgrounds. By improving retention rates this maintains diversity and inclusivity in academia

In conclusion, it is important for an educational institution to understand and observe their student drop out and graduation rates as it influences a student’s success, quality of education, compliance and ranking of the institution. With a better understanding of what factors influence a student to graduate or dropout, they can work towards more efficient strategies that improve and expand their programs, finances and the communities they serve.

References: Attribute Value Tables

A screenshot of a computer

Description automatically generatedA screenshot of a screen

Description automatically generated

A list of words on a white background

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A list of jobs

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated