Learning Analytics Final Project Part 1 - Data Prep

Maria Baldridge

May 5, 2023

1: Overview

The purpose of my project is to identify certain characteristics of successful learners using personal, educational, and behavioral information. This information would be used to predict successful outcomes, specifically, with the goal of identifying how much support students would need. This information would further be used to determine habits and choices that increase the likelihood of success in a college course, which could then be used to develop academic plans or learning tools/methods for students.

Questions

My key question is:

"Can I use personal/demographic data to predict student success in a college course?"

The follow up question is:

"Can I use the behavioral data to determine what habits/actions foster student success?"

Audience

The audience would be both teachers and students. My key question would be teacher-facing and my follow up question would be both teacher (to develop any academic success plans) and student-facing (tips and advice on how to do well in their courses).

How is the question and plan related to learning?

My questions are related to learning as they would use data collected from students in these courses in order to improve student learning outcomes. My key question, being able to predict student success, would be used to allow instructors some foresight on which students might need more support in their classes. My follow up question would be used to provide general support in the form of learning tips/tools for students on how they can increase their chances of success and for instructors to formulate academic plans that might help students at risk of failing courses.

Usefulness and Potential Actions

This data could be useful in determining what characteristics and habits are conducive to success in the college setting. Both students and teachers could use this information to improve learning outcomes by adopting/encouraging these habits. Instructors could use the information to get an idea on how the students and class cohort might do in a given course and revise their lessons/lectures to meet the needs of the students.

2: Describe Collection, Acquisition, and Storage of Data

The data was collected by two researchers, Nevriye Yilmaz and Boran Sekeroglu, of Yakın Doğu Üniversitesi (Near East University), Turkey. The data was collected using questionnaires completed by

students within the Faculty of Engineering and Faculty of Educational Sciences in 2019 (Yilmaz & Sekeroglu, 2021).

I acquired this dataset after an online search led me to UC Irvine's Machine Learning Repository. I searched for datasets on the topic of "learning" and found this dataset by Yilmaz and Sekeroglu. The datasets are anonymized and have been shared on public platforms. Since the data is anonymized and public, I am storing it on my laptop computer without any additional security measures (Yilmaz & Sekeroglu, 2021).

3: Understand Your Dataset

Variables and Definitions

Variables	What they represent
Student ID	Numerical ID of each student
Student Age	Age of the student in years
Sex	Whether the student is male or female. Only a binary choice was recorded.
High School Type	Whether the student graduate from a public, private, or other high school.
Scholarship Type	Whether the student got 0%, 25%, 50%, 75%, 100% scholarship awarded.
Work	Whether the student works (Yes/No).
Artistic/Sports Activity	Whether the student participates in any regular artistic or sports activity (Yes/No).
Partner	Whether the student has a significant other (Yes/No).
Salary	Student's salary in USD (Range).
Transportation	How the student get to college.
Housing	What type of housing the student lives in.
Mother's Education	The highest level of school that the students mother attended.
Father's Education	The highest level of school that the student's father attended.
# of Siblings	The number of siblings the student has.
Parents Status	Parent's marital status
Mother's Occupation	The general field of occupation for the mother.
Father's Occupation	The general field of occupation for the father.
Weekly Study Hours	The number of hours the student studies per week (Range).
Reading Frequency (non-scientific)	The frequency that the student reads non-scientific books and journals.
Reading Frequency (scientific)	The frequency that the student reads scientific books and journals.
Attend. Seminars/Conferences Related to Department	Whether the student attends any conferences or seminars related to their department (Yes/No).

Impact of Projects/Activities on Success	What kind of impact projects and activities had on the student's success in the course.
Attendance	How often the students attended classes.
Preparation to Midterm Exams 1	Whether the students studied for these exams alone, with friends, or not applicable.
Preparation to Midterm Exams 2	Whether the students studied for these exams close to the date, regularly throughout the semester, or never.
Taking Notes in Classes	Whether the students took notes during class.
Listening in Classes	Whether the student listened in class.
Discussion improves my interest and success in the course	Whether discussions improved the student's interest and outcome in the course.
Flip-Classroom	Whether flip-classroom model was not useful, useful, or not applicable.
Cumulative GPA in Last Semester (/4.0)	The student's cumulative GPA in the last semester (Range).
Expected Cumulative GPA at Graduation (/4.0)	What the student expects their GPA to be at graduation (Range).
Course ID	Which course(s) the student was enrolled in.
Output Grade	The grade the student earned in the course (Letter Range).

Signal/Noise Identification

Variables	Signal or Noise?
Student ID	Signal
Student Age	Signal
Sex	Signal
High School Type	Signal
Scholarship Type	Signal
Work	Signal
Artistic/Sports Activity	Noise
Partner	Noise
Salary	Signal
Transportation	Noise
Housing	Signal
Mother's Education	Noise
Father's Education	Noise
# of Siblings	Noise
Parents Status	Noise
Mother's Occupation	Noise
Father's Occupation	Noise
Weekly Study Hours	Signal
Reading Frequency (non-scientific)	Signal

Reading Frequency (scientific)	Signal
Attend. Seminars/Conferences Related to Department	Signal
Impact of Projects/Activities on Success	Signal
Attendance	Signal
Preparation to Midterm Exams 1	Signal
Preparation to Midterm Exams 2	Signal
Taking Notes in Classes	Signal
Listening in Classes	Signal
Discussion improves my interest and success in the course	Signal
Flip-Classroom	Noise
Cumulative GPA in Last Semester (/4.0)	Signal
Expected Cumulative GPA at Graduation (/4.0)	Signal
Course ID	Signal
Output Grade	Signal

Will you need to remove any variables, filter anything out, or work with missing data?

I think removing variables would only be necessary for convenience. I did not need to filter anything or work with missing data, but I did need to wrangle some of the data that was reported in ranges in order to analyze it quantitatively.

Do you have enough data and the right data to address your questions?

Yes, I believe that I have enough and the right kind of data to address my questions.

4. Data Cleaning Sub-cycle - Refining Your Questions and Plan

Going through this process helped me to refine my question by showing me that there was just so much data that I could possibly include. Narrowing down my variables was necessary due to the nature of how the data was reported and the limitations of how it could be displayed in Tableau. The cleaning sub-cycle also highlighted that while some of this information is available because the students reported it in the questionnaire, in a regular college course an instructor/professor would not have access to some of this information (relationship status, parental employment, etc.), rendering it irrelevant to my analysis.

5. Data Cleaning Sub-cycle - Steps for Wrangling

- The data was in .csv format and could be directly imported into Tableau.
- I renamed all the variable columns because they were only labeled with numbers (“Column 1”).
- The data was coded so that numbers represented every data value (e.g., for the Sex section the response was 1 for female and 2 for male). Since Tableau read the data entries as numerical, the data was imported as quantitative measurements. To create meaningful charts of the data I needed to convert the values from “measure” to “dimension.” To make the values readable in

charts (rather than numerical coding), I had to give each new dimension an alias that matched the attribute information provided with the .csv file.

- Much of the data was either qualitative (yes/no) and data that could have been quantitative (GPA, age, etc.) was binned in the data file. For the data points that could be analyzed quantitatively (GPA) but were presented in bins, I created a calculated column that approximated each bin as the median value of the bin and applied it to the data. For example, the bin for a GPA of 2.5-2.99 was coded as a "3", so I used the formula: (value+2.5)/2 which yields a numerical median GPA of 2.75. This is not precise, but it allowed for the GPA data to be analyzed numerically while keeping the data within the parameters.

6. Reflect on Ethics

The dataset was already anonymized and available publicly, so I have no concerns regarding confidentiality or security of storage. The ethics concerns that I have that are related to this analysis stem from its possible use. In trying to use characteristics of students to predict success, there is a potential for discrimination, issues of equity, and the introduction of another source of bias for the students. Depending on how the data is used and the character of the person using them, I can imagine some problematic uses in how this information could be used, such as to exclude students from programs or give less support because of a fatalistic mindset from teachers/admin. This same fatalistic mindset in students could also create a sort of self-fulfilling issue where a student would find out they did not have the same personal characteristics that successful students have, creating a mental barrier that the student would then have to overcome.

References

YÄ±lmaz N., Sekeroglu B. (2020) Student Performance Classification Using Artificial Intelligence Techniques. In: Aliev R., Kacprzyk J., Pedrycz W., Jamshidi M., Babanli M., Sadikoglu F. (eds) 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions - ICSCCW-2019. ICSCCW 2019. Advances in Intelligent Systems and Computing, vol 1095. Springer, Cham.